Г.И.Просветов

ЭКОНОМЕТРИКА

ЗАДАЧИ И РЕШЕНИЯ

Учебно-методическое пособие

РДЛ

Г.И.Просветов

ЭКОНОМЕТРИКА

ЗАДАЧИ И РЕШЕНИЯ

Учебно-методическое пособие

2-е издание

МОСКВА



2005

ББК 65в6я73
П82

Рецензенты: В. В. ШЕМЕТОВ,
д.э.н., профессор, заведующий кафедрой менеджмента организации и маркетинга Российской академии предпринимательства

В. Л. МИРОНОВ,
к.ф.-м.н., доцент Института бизнеса и делового администрирования Академии народного хозяйства при Правительстве Российской Федерации

Просветов Г. И.

П82 **Эконометрика. Задачи и решения:** Учебно-методическое пособие. 2-е изд. — М.: Издательство РДЛ, 2005. — 104 с.

ISBN 5-93840-079-1

В настоящем пособии на простых примерах раскрываются такие разделы эконометрики, как парная и множественная регрессии (метод наименьших квадратов (МНК) и его предпосылки, оценка линейности связи, коэффициенты корреляции и детерминации, анализ статистической значимости коэффициентов, доверительные интервалы и испытания гипотез в линейном регрессионном анализе, тест Чоу), гетероскедастичность, автокорреляция, мультиколлинеарность, фиктивные переменные, нелинейные связи, порядковые испытания, временные ряды (аддитивные и мультипликативные модели, построение прогноза), экспоненциальное сглаживание (простая модель и с поправкой на тренд), системы одновременных уравнений (косвенный МНК, проблема идентификации, двухшаговый МНК). Также рассмотрены такие важные темы математической статистики, как доверительные интервалы и испытания гипотез.

Пособие содержит программу курса, задачи для самостоятельного решения с ответами и задачи для контрольной работы. Издание рассчитано на преподавателей и студентов экономических специальностей высших учебных заведений.

ББК 65в6я73

ISBN 5-93840-079-1

Содержание

ООО «Издательство РДЛ».
117334, Москва, ул. Вавилова, д. 30/6.
Тел.: (095) 135-98-93. e-mail: rdl@rinet.ru

Лицензия ИД № 00834 от 25 января 2000 г.

ISBN 5-93840-079-1



Сдано в набор 8.12.2004.
Подписано в печать 20.02.2005.
Формат 84x108 1/32. Гарнитура Школьная.
Печать офсетная. Тираж 700 экз.
Заказ № 339.

Начальник редакции *В. М. Дубильт.*
Научный редактор *В. М. Трояновский.*

Отпечатано в Загорской типографии
141300, Московская область,
г. Сергиев Посад, пр. Красной Армии, д. 212Б.

Предисловие

Эконометрика — это наука, в которой на базе реальных статистических данных строятся, анализируются и совершенствуются математические модели реальных экономических явлений. Одним из важнейших направлений эконометрики является построение прогнозов по различным экономическим показателям. В настоящее время имеется ряд обстоятельных руководств по эконометрике. Но, по мнению автора, всем им присущ один существенный недостаток — они не учитывают реальные учебные планы обучения студентов экономических специальностей вузов. Предлагаемое пособие знакомит читателя с рядом важнейших разделов эконометрики и призвано помочь тем, кто осваивает этот курс, особенно в системе заочного и вечернего образования.

Традиционно предполагается, что студенты, изучающие эконометрику, уже прослушали курс теории вероятностей и математической статистики. На практике же большинство из них не обладает требуемым уровнем математической и статистической подготовки. Поэтому в первых двух главах рассмотрены такие важные разделы математической статистики, как построение доверительных интервалов и испытание гипотез.

Третья и четвертая главы посвящены вопросам линейной регрессии (соответственно парной и множественной). В них представлен фундаментальный метод оценки параметров уравнений регрессии — метод наименьших квадратов.

Пятая, шестая и седьмая главы затрагивают проблему невыполнимости предпосылок метода наименьших квадратов (гетероскедастичность, автокорреляция, мультиколлинеарность). В них приводятся способы обнаружения и смягчения последствий.

В восьмой главе изучаются модели, содержащие вместе с количественными переменными и качественные переменные (фиктивные переменные). В девятой главе рассмотрены нелинейные регрессионные модели. Тема десятой главы — порядковые испытания.

В одиннадцатой и двенадцатой главах изучаются основные понятия временных рядов, способы построения прогнозов, метод скользящей средней и экспоненциальное сглаживание.

В тринадцатой главе анализируются системы одновременных уравнений, рассматриваются методы нахождения оценок для таких систем и исследуется проблема идентификации.

Весь материал пособия разбит на главы, а главы — на параграфы. Каждый параграф — это отдельная тема курса «Эконометрика». В начале параграфа приводится необходимый минимум теоретических сведений, затем подробно разбираются модельные примеры. После каждого разобранного примера приводится задача для самостоятельного решения. Ответы ко всем задачам помещены в конце книги. Пособие содержит также программу курса и задачи для контрольной работы.

В книге показано, как можно избежать долгих и утомительных вычислений с помощью пакета Excel (надстройка «Пакет анализа» и встроенные статистические функции). Обычно книги по эконометрике содержат большое количество табулированных значений всевозможных распределений. Автор сознательно отказался от этого, заменив ссылками на соответствующие статистические функции пакета Excel.

За основу книги приняты курсы, читаемые автором в Российской академии предпринимательства. Всем студентам, прослушавшим эти курсы, автор выражает благодарность за продуктивную совместную работу.

Автор выражает искреннюю признательность В. М. Трояновскому за многочисленные замечания, способствовавшие улучшению книги.

Автор

Глава 1

ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ

Изучаемая генеральная совокупность может быть очень большой. Поэтому с целью экономии времени и материальных ресурсов случайным образом производят выборку из генеральной совокупности. Для этой выборки вычисляют выборочную среднюю $\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n}$, выборочную дисперсию $s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n}$ и интересующие нас параметры. Как оценить параметры генеральной совокупности, зная эти параметры для выборки?

Для генеральной совокупности строится *доверительный интервал* — интервал значений, в пределах которого, как мы можем надеяться, находится параметр генеральной совокупности. Наша надежда выражается *доверительной вероятностью* — вероятностью, с которой доверительный интервал «захватит» истинное значение параметра генеральной совокупности. Чем выше доверительная вероятность, тем шире доверительный интервал. Значение доверительной вероятности выбирает сам исследователь. Обычно это 0,9; 0,95; 0,99.

§ 1.1. ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ГЕНЕРАЛЬНОЙ СРЕДНЕЙ a (генеральная дисперсия σ^2 известна)

Если генеральная совокупность подчиняется нормальному закону распределения с известной дисперсией σ^2, то

$$\overline{X} - z_{\alpha/2}\sigma/\sqrt{n} < a < \overline{X} + z_{\alpha/2}\sigma/\sqrt{n},$$

где $\overline{X}$ — выборочная средняя, n — объем выборки, $\alpha = 1 - p$, p — доверительная вероятность, $z_{\alpha/2}$ берем из таблицы.

α	0,4	0,25	0,2	0,15	0,1	0,05	0,025	0,01	0,005	0,001
z_α	0,253	0,675	0,842	1,036	1,282	1,645	1,960	2,326	2,576	3,090

Для вычисления $z_{\alpha/2}$ можно также воспользоваться статистической функцией НОРМСТОБР(1 − α) мастера функций f_x пакета Excel.

Пример 1. Автомат, работающий со стандартным отклонением $\sigma = 5$ г, фасует чай в пачки. Проведена случайная выборка объемом $n = 30$ пачек. Средний вес пачки чая в выборке $\overline{X} = 101$ г. Найдем доверительный интервал для среднего веса пачки чая в генеральной совокупности с доверительной вероятностью $p = 95\%$.

$p = 0{,}95 \Rightarrow \alpha = 1 - p = 1 - 0{,}95 = 0{,}05 \Rightarrow \alpha/2 = 0{,}025 \Rightarrow z_{\alpha/2} = 1{,}96$.

$\overline{X} \pm z_{\alpha/2}\sigma/\sqrt{n} = 101 \pm 1{,}96 \times 5/\sqrt{30} \approx 101 \pm 1{,}79$, то есть искомый интервал (99,21; 102,79).

Задача 1. Автомат, работающий со стандартным отклонением $\sigma = 3$ г, фасует чай в пачки. Проведена случайная выборка объемом $n = 40$ пачек. Средний вес пачки чая в выборке $\overline{X} = 79$ г. Найти доверительный интервал для среднего веса пачки чая в генеральной совокупности с доверительной вероятностью $p = 99\%$.

Замечание. Вместо вычислений по формуле $z_{\alpha/2}\sigma/\sqrt{n}$ можно было бы воспользоваться функцией ДОВЕРИТ(α; σ; n) мастера функций f_x пакета Excel.

§ 1.1.1. Объем выборки, необходимый для оценки генеральной средней

Пример 2. Вернемся к примеру 1. Мы получили доверительный интервал $\overline{X} \pm z_{\alpha/2}\sigma/\sqrt{n} \approx 101 \pm 1{,}79$. Предположим, что нам нужна ширина доверительного интервала ± 1 грамм. Каким должен быть объем выборки?

$z_{\alpha/2}\sigma/\sqrt{n} \le 1 \Rightarrow \sqrt{n} \ge z_{\alpha/2}\sigma \Rightarrow n \ge (z_{\alpha/2}\sigma)^2 = (1{,}96 \times 5)^2 = 96{,}04$, то есть минимальный объем выборки равен 97. Так как объем первоначальной выборки равен 30, то объем новой выборки равен $97 - 30 = 67$ пачек. Находим среднюю $\overline{X}$ для объединенной выборки в 97 пачек (находим именно среднюю для выборки в 97 единиц, а не среднее арифметическое средних для выборок объемов 30 и 67 пачек) и получаем доверительный интервал для средней в генеральной совокупности $\overline{X} \pm 1$.

Задача 2. Каким должен быть объем выборки в задаче 1, если требуемая ширина доверительного интервала $\pm 0{,}5$ г?

§ 1.2. ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ГЕНЕРАЛЬНОЙ СРЕДНЕЙ a (генеральная дисперсия σ^2 неизвестна)

Если генеральная совокупность подчиняется нормальному закону распределения с неизвестной дисперсией σ^2, то

$$\overline{X} - t_{\alpha/2,n-1}s/\sqrt{n-1} < a < \overline{X} + t_{\alpha/2,n-1}s/\sqrt{n-1}\ ,$$

где $\overline{X}$ — выборочная средняя, n — объем выборки, $\alpha = 1 - p$, p — доверительная вероятность, s — выборочное стандартное отклонение. Значение $t_{\alpha/2,n-1}$ берем из таблицы t-распределения (распределения Стьюдента). Для вычисления значения $t_{\alpha/2,n-1}$ также можно воспользоваться статистической функцией СТЬЮДРАСПОБР (α; $n-1$) мастера функций f_x пакета Excel.

Пример 3. Автомат фасует чай в пачки. Проведена случайная выборка объемом n = 30 пачек. Средний вес пачки чая в выборке $\overline{X}$ = 101 г, выборочное стандартное отклонение s = 4 г. Найдем доверительный интервал для среднего веса пачки чая в генеральной совокупности с доверительной вероятностью p = 95%.

$p = 0{,}95 \Rightarrow \alpha = 1 - p = 1 - 0{,}95 = 0{,}05 \Rightarrow \alpha/2 = 0{,}025.$
$n = 30 \Rightarrow n - 1 = 29 \Rightarrow t_{\alpha/2,n-1} = t_{0{,}025;\,29} \approx 2{,}045.$
$\overline{X} \pm t_{\alpha/2,n-1}s/\sqrt{n-1} \approx 101 \pm 2{,}045 \times 4/\sqrt{29} \approx 101 \pm 1{,}52$, то есть искомый интервал (99,48; 102,52).

Задача 3. Автомат фасует чай в пачки. Проведена случайная выборка объемом n = 46 пачек. Средний вес пачки чая в выборке $\overline{X}$ = 79 г, выборочное стандартное отклонение s = 3 г. Найти доверительный интервал для среднего веса пачки чая в генеральной совокупности с доверительной вероятностью p = 99%.

§ 1.2.1. Объем выборки, необходимый для оценки генеральной средней

Пример 4. Вернемся к примеру 3. Мы получили доверительный интервал $\overline{X} \pm t_{\alpha/2,n-1}s/\sqrt{n-1} \approx 101 \pm 1{,}52$. Предположим, что нам нужна ширина доверительного интервала ± 1 г. Каким должен быть тогда объем выборки?

$t_{\alpha/2,n-1}s/\sqrt{n-1} \le 1 \Rightarrow \sqrt{n-1} \ge t_{\alpha/2,n-1}s \Rightarrow n-1 \ge (t_{\alpha/2,n-1}s)^2$ $\Rightarrow n \ge 1+(t_{\alpha/2,n-1}s)^2 \approx 1+(2{,}045\times 4)^2 \approx 67{,}9$, то есть минимальный объем выборки равен 68. Но плохо то, что $t_{\alpha/2,n-1}$ зависит от n. Тем не менее, полученный результат можно использовать. На самом деле n будет меньше.

Если полученное значение $n \ge 30$, то можно вместо $t_{\alpha/2,n-1}$ рассмотреть $z_{\alpha/2}$ и воспользоваться формулой $z_{\alpha/2}s/\sqrt{n-1} \le 1 \Rightarrow n \ge 1+(z_{\alpha/2}s)^2 \approx 1+(1{,}96\times 4)^2 \approx 62{,}47$, то есть минимальный объем выборки равен 63. Так как объем первоначальной выборки равен 30, то объем новой выборки равен $63-30=33$ пачки. Находим среднюю $\overline{X}$ для объединенной выборки в 63 пачки и получаем доверительный интервал для средней в генеральной совокупности $\overline{X} \pm 1$.

Задача 4. Каким должен быть объем выборки в задаче 3, если требуемая ширина доверительного интервала $\pm 0{,}5$ г?

§ 1.3. ДОВЕРИТЕЛЬНЫЙ ИНТЕРВАЛ ДЛЯ ГЕНЕРАЛЬНОЙ ДОЛИ

Очень часто нас интересует, какова *генеральная доля* — доля объектов генеральной совокупности, обладающих определенным свойством.

Производится выборка объема n. Для нее вычисляется *выборочная доля* $\hat{p}$ — доля объектов, обладающих этим свойством. Тогда при выполнении условий $n\hat{p} \ge 5$, $n(1-\hat{p}) \ge 5$ доверительный интервал для генеральной доли задается формулой $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$.

Пример 5. Проведена выборка объема $n = 2000$ шт. 150 из них оказались бракованными. Найдем доверительный интервал доли бракованных изделий в генеральной совокупности для доверительной вероятности $p = 95\%$.

$\hat{p} = 150/2000 = 0{,}075$.

$n\hat{p} = 2000\times 0{,}075 = 150 > 5$.

$n(1-\hat{p}) = 2000\times(1-0{,}075) = 1850 > 5$.

Оба условия выполнены.

$p = 0{,}95 \Rightarrow \alpha = 1-p = 1-0{,}95 = 0{,}05 \Rightarrow \alpha/2 = 0{,}025 \Rightarrow$ $\Rightarrow z_{\alpha/2} = 1{,}96$.

$\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} = 0{,}075 \pm 1{,}96\sqrt{0{,}075(1-0{,}075)/2000} \approx$ $\approx 0{,}075 \pm 0{,}012$.

То есть искомый интервал (0,063; 0,087).

Задача 5. Проведена выборка объема $n = 1000$ шт. 120 из них оказались бракованными. Найти доверительный интервал доли бракованных изделий в генеральной совокупности для доверительной вероятности $p = 99\%$.

§ 1.3.1. Объем выборки, необходимый для оценки генеральной доли

Пример 6. Вернемся к примеру 5. Мы получили доверительный интервал $\hat{p} \pm z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \approx 0{,}075 \pm 0{,}012$. Предположим, что нам нужна ширина доверительного интервала $\pm 0{,}005$. Каким должен быть тогда объем выборки?

$$z_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n} \leq 0{,}005 \Rightarrow$$
$$\Rightarrow (z_{\alpha/2})^2\hat{p}(1-\hat{p})/n \leq 0{,}005^2 = 0{,}000025 \Rightarrow$$
$$\Rightarrow n \geq (z_{\alpha/2})^2\hat{p}(1-\hat{p})/0{,}000025 \approx$$
$$\approx 1{,}96^2 \times 0{,}075 \times (1-0{,}075)/0{,}000025 \approx 10660.$$

То есть минимальный объем выборки равен 10660. Так как объем первоначальной выборки равен 2000, то объем новой выборки равен $10660 - 2000 = 8660$ деталей. Находим выборочную долю бракованных изделий $\hat{p}$ для объединенной выборки в 8660 деталей и получаем доверительный интервал для доли бракованных изделий в генеральной совокупности $\hat{p} \pm 0{,}005$.

Задача 6. Каким должен быть объем выборки в задаче 5, если требуемая ширина доверительного интервала $\pm 0{,}003$?

Глава 2

ИСПЫТАНИЕ ГИПОТЕЗ

§ 2.1. ПРОЦЕДУРА ИСПЫТАНИЯ ГИПОТЕЗ

Очень часто генеральная совокупность должна подчиняться некоторым параметрам. Например, фасовочная машина должна наполнять пакеты сахаром по 1 кг. Как узнать, действительно ли генеральная совокупность подчиняется этим ограничениям? С этой целью проводят *испытание гипотез.*

Из генеральной совокупности проводят выборку объема n. Для этой выборки вычисляют нужные характеристики. Затем формулируют две гипотезы: основную H_0 и альтернативную H_1. Основная гипотеза H_0 — это то утверждение, которое подлежит проверке.

Например, гипотеза H_0: генеральная средняя $a = 2$. Альтернативная гипотеза H_1 в этом примере может быть сформулирована любым из следующих трех способов:

а) H_1: $a > 2$ (правосторонняя проверка);
б) H_1: $a < 2$ (левосторонняя проверка);
в) H_1: $a \neq 2$ (двусторонняя проверка).

Исследователь задает доверительную вероятность p — величину, которая отражает степень уверенности исследователя в результате испытания. Для односторонней проверки $\alpha = 1 - p$, для двусторонней проверки $\alpha = (1 - p)/2$. Величина $1 - p$ называется *уровнем значимости.*

По α, n в зависимости от вида решаемой задачи по таблицам находят одну (для односторонней проверки) или две (для двусторонней проверки) граничные точки, которые наносят на координатную ось. Порядок нахождения граничных точек показан далее.

По результатам выборки вычисляется величина, называемая *статистикой*. Формула для вычисления статистики зависит от вида решаемой задачи. Значение статистики наносят на координатную ось. В зависимости от взаимного расположения значения статистики и граничных точек возможен один из трех вариантов:

1) принимается H_0;

2) отклоняется H_0 и без всякой проверки принимается H_1;

3) доказательство является неубедительным, нужно больше данных.

Для левосторонней проверки:

Отклонение H_0 Принятие H_1 $(1-p)\%$	Принятие H_0 $p\%$

граничная точка

Для правосторонней проверки:

Принятие H_0 $p\%$	Отклонение H_0 Принятие H_1 $(1-p)\%$

граничная точка

Для двусторонней проверки:

Отклонение H_0 Принятие H_1 $[(1-p)/2]\%$	Принятие H_0 $p\%$	Отклонение H_0 Принятие H_1 $[(1-p)/2]\%$

граничная точка граничная точка

Чем выше доверительная вероятность, тем шире область принятия H_0.

§ 2.2. ИСПЫТАНИЕ ГИПОТЕЗ НА ОСНОВЕ ВЫБОРОЧНОЙ СРЕДНЕЙ ПРИ ИЗВЕСТНОЙ ГЕНЕРАЛЬНОЙ ДИСПЕРСИИ σ^2

Для выборки объема n вычисляется выборочная средняя $\overline{X}$. a — предполагаемое значение генеральной средней. Граничные точки: z_α (для правосторонней проверки), $-z_\alpha$ (для левосторонней проверки), $\pm z_\alpha$ (для двусторонней проверки). Значение z_α находим по таблице (*см. § 1.1*). Статистика

$$z = \frac{\overline{X} - a}{\sigma/\sqrt{n}}.$$

Пример 7. Автомат, работающий со стандартным отклонением $\sigma = 1$ г, фасует чай в пачки со средним весом $a = 100$ г. В случайной выборке объема $n = 25$ пачек средний вес $\overline{X} = 101{,}5$ г. Надо ли отрегулировать автомат? Доверительная вероятность $p = 95\%$.

H_0: для нормальной совокупности генеральная средняя $a = 100$ г.

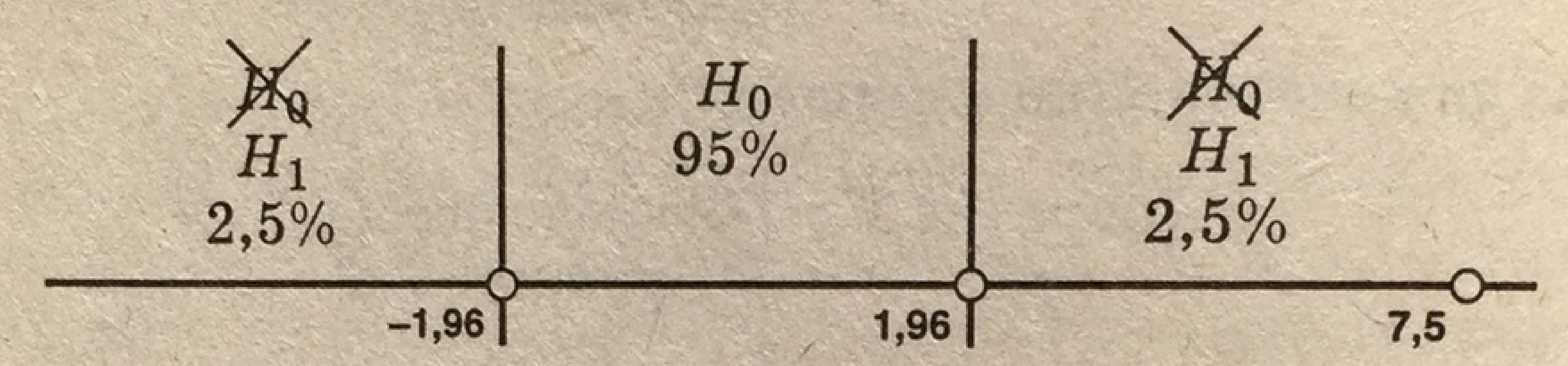
H_1: $a \neq 100$ г.

Проведем двустороннюю проверку.

$\alpha = (1-p)/2 = (1-0,95)/2 = 0,025 \Rightarrow z_\alpha = 1,96 \Rightarrow$ граничные точки $\pm 1,96$.

Статистика $z = \dfrac{\overline{X} - a}{\sigma/\sqrt{n}} = \dfrac{101,5 - 100}{1/\sqrt{25}} = 7,5.$

Отметим значения на числовой оси.



Отклоняем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Автомат нужно отрегулировать.

Задача 7. Автомат, работающий со стандартным отклонением $\sigma = 1,5$ г, фасует чай в пачки со средним весом $a = 80$ г. В случайной выборке объема $n = 16$ пачек средний вес $\overline{X} = 78,5$ г. Надо ли отрегулировать автомат? Доверительная вероятность $p = 99\%$.

Пример 8. Станок, работающий со стандартным отклонением $\sigma = 0,5$ мм, производит детали средней длины $a = 20$ мм. В случайной выборке объема $n = 16$ деталей средняя длина $\overline{X} = 19,8$ мм. Правильно ли настроен станок? Доверительная вероятность $p = 99\%$.

H_0: для нормальной совокупности генеральная средняя $a = 20$ мм.

H_1: $a < 20$ мм.

Проведем левостороннюю проверку.

$\alpha = 1 - p = 1 - 0,99 = 0,01 \Rightarrow z_\alpha = 2,326 \Rightarrow$ граничная точка $-2,326$.

Статистика $z = \dfrac{\overline{X} - a}{\sigma/\sqrt{n}} = \dfrac{19,8 - 20}{0,5/\sqrt{16}} = -1,6.$

Отметим значения на числовой оси.

H_0 H_1 1% | H_0 99%

–2,326   –1,6

Принимаем гипотезу H_0 на уровне значимости 1%. Станок настроен правильно.

Задача 8. Станок, работающий со стандартным отклонением $\sigma = 0{,}4$ мм, производит детали средней длины $a = 30$ мм. В случайной выборке объема $n = 25$ деталей средняя длина $\overline{X} = 30{,}1$ мм. Правильно ли настроен станок? Доверительная вероятность $p = 95\%$.

§ 2.3. ИСПЫТАНИЕ ГИПОТЕЗ НА ОСНОВЕ ВЫБОРОЧНОЙ СРЕДНЕЙ ПРИ НЕИЗВЕСТНОЙ ГЕНЕРАЛЬНОЙ ДИСПЕРСИИ

Для выборки объема n вычисляются выборочная средняя $\overline{X}$ и выборочное стандартное отклонение s. Пусть a — предполагаемое значение генеральной средней. По таблице t-распределения находим $t_{\alpha,n-1}$. В Excel для двусторонней проверки $t_{\alpha,n-1}$ = СТЬЮДРАСПОБР($1-p$; $n-1$), для односторонней проверки $t_{\alpha,n-1}$ = СТЬЮДРАСПОБР($2(1-p)$; $n-1$). Граничные точки: $t_{\alpha,n-1}$ (для правосторонней проверки), $-t_{\alpha,n-1}$ (для левосторонней проверки), $\pm t_{\alpha,n-1}$ (для двусторонней проверки). Статистика $t = \dfrac{\overline{X}-a}{s/\sqrt{n-1}}$.

Пример 9. Производитель утверждает, что средний вес пачки чая не меньше $a = 100$ г. Инспектор отобрал 10 пачек чая и взвесил. Их вес оказался 97, 102, 103, 98, 96, 105, 98, 100, 101 и 99 г соответственно. Не противоречит ли это утверждению производителя? Предполагается, что вес пачек чая распределен нормально. Доверительная вероятность $p = 99\%$.

Номер пачки	Вес, г x_i	$x_i-\overline{X}$	$(x_i-\overline{X})^2$
1	97	−2,9	8,41
2	102	2,1	4,41
3	103	3,1	9,61
4	98	−1,9	3,61
5	96	−3,9	15,21
6	105	5,1	26,01
7	98	−1,9	3,61
8	100	0,1	0,01
9	101	1,1	1,21
10	99	−0,9	0,81
Сумма	999	0	72,9

H_0: для нормальной совокупности генеральная средняя $a = 100$ г.

H_1: $a < 100$ г.

Проведем левостороннюю проверку.

$\alpha = 1 - p = 1 - 0{,}99 = 0{,}01 \Rightarrow t_{\alpha,n-1} = t_{0{,}01;\,10-1} = 2{,}821 \Rightarrow$ граничная точка $-2{,}821$.

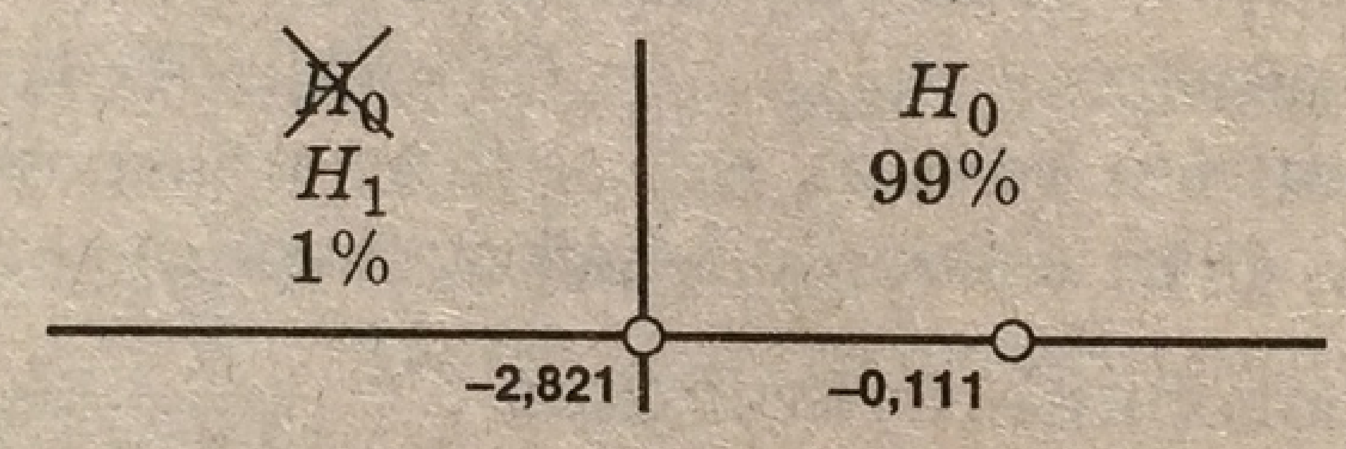
Найдем $\overline{X}$ и s.

$$\overline{X} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{999}{10} = 99{,}9 \text{ г.}$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n} = \frac{72{,}9}{10} = 7{,}29 \text{ г}^2. \quad s = \sqrt{7{,}29} = 2{,}7 \text{ г.}$$

Статистика $t = \dfrac{\overline{X} - a}{s/\sqrt{n-1}} = \dfrac{99{,}9 - 100}{2{,}7/\sqrt{10-1}} \approx -0{,}111.$

Отметим значения на числовой оси.



Принимаем гипотезу H_0 на уровне значимости 1%. Выборка инспектора не противоречит утверждению производителя.

Задача 9. Производитель утверждает, что средний вес плитки шоколада не меньше $a = 50$ гр. Инспектор отобрал 10 пачек чая и взвесил. Их вес оказался 49, 50, 51, 52, 48, 47, 49, 52, 48 и 51 г соответственно. Не противоречит ли это утверждению производителя? Предполагается, что вес плитки шоколада распределен нормально. Доверительная вероятность $p = 95\%$.

§ 2.4. ИСПЫТАНИЕ ГИПОТЕЗ НА ОСНОВЕ ВЫБОРОЧНОЙ ДОЛИ

Для выборки объема n вычисляется выборочная доля $\hat{p}$ = (число элементов выборки, обладающих нужным свойством)/(объем выборки) и сравнивается с генеральной долей $\overline{p}$. Граничные точки: z_α (для правосторонней проверки), $-z_\alpha$ (для левосторонней проверки), $\pm z_\alpha$ (для двусторонней проверки). Значение z_α находим по таблице (*см. § 1.1*).

Статистика $z = \dfrac{\hat{p} - \overline{p}}{\sqrt{\overline{p}(1-\overline{p})/n}}$.

Пример 10. Производитель утверждает, что доля бракованных изделий не превосходит 3%. В случайной выборке объема $n = 100$ изделий оказалось 5 бракованных изделий. Не противоречит ли это утверждению производителя? Доверительная вероятность $p = 95\%$.

H_0: доля бракованных изделий равна 3%, то есть $\overline{p} = 0{,}03$.

H_1: $\overline{p} > 0{,}03$.

Проведем правостороннюю проверку.

$\alpha = 1 - p = 1 - 0{,}95 = 0{,}05 \Rightarrow z_\alpha = 1{,}645$. Это граничная точка.

Оценка $\hat{p} = 5/100 = 0{,}05$.

Статистика $z = \dfrac{\hat{p} - \overline{p}}{\sqrt{\overline{p}(1-\overline{p})/n}} = \dfrac{0{,}05 - 0{,}03}{\sqrt{0{,}03(1-0{,}03)/100}} \approx 1{,}172.$

Отметим значения на числовой оси.

H_0 95% | ~~H_0~~ H_1 5%

1,172 | 1,645

Принимаем гипотезу H_0 на уровне значимости 5%. Выборка не противоречит утверждению производителя.

Задача 10. Производитель утверждает, что доля бракованных изделий не превосходит 7%. В случайной выборке объема $n = 150$ изделий оказалось 16 бракованных изделий. Не противоречит ли это утверждению производителя? Доверительная вероятность $p = 99\%$.

§ 2.5. ИСПЫТАНИЕ ГИПОТЕЗ О ДВУХ ГЕНЕРАЛЬНЫХ ДИСПЕРСИЯХ

Очень часто про две независимые выборки объема n_1 и n_2 соответственно нужно узнать, взяты ли они из нормальных генеральных совокупностей с одинаковой дисперсией. Для каждой выборки находим выборочную дисперсию s_1^2 и s_2^2 соответственно. Оценка генеральной дисперсии по первой выборке $\sigma_1^2 = n_1 s_1^2/(n_1 - 1)$. Оценка генеральной дисперсии по второй выборке $\sigma_2^2 = n_2 s_2^2/(n_2 - 1)$. Статисти-

ка F = (бо́льшая оценка генеральной дисперсии)/(меньшая оценка генеральной дисперсии).

Обозначим через n_A объем выборки, у которой больше оценка генеральной дисперсии, через n_B обозначим объем другой выборки. Так как дисперсия неотрицательна, то нам потребуется одна граничная точка $F_{\alpha;n_A-1;n_B-1}$, которую находят из таблицы F-распределения (распределения Фишера). Можно воспользоваться статистической функцией FРАСПОБР (α; n_A-1; n_B-1) мастера функций f_x пакета Excel.

Пример 11. Инвестиция 1 рассчитана на $n_1 = 12$ лет, дисперсия ежегодных прибылей $s_1^2 = (20\%)^2$. Инвестиция 2 рассчитана на $n_2 = 10$ лет, дисперсия ежегодных прибылей $s_2^2 = (30\%)^2$. Предполагается, что распределение ежегодных прибылей на инвестиции подчиняется нормальному закону распределения. Равны ли риски инвестиций 1 и 2? Доверительная вероятность $p = 95\%$.

H_0: $\sigma_1^2 = \sigma_2^2$.

H_1: $\sigma_1^2 \neq \sigma_2^2$.

Оценка генеральной дисперсии по первой выборке $\sigma_1^2 = n_1 s_1^2/(n_1-1) = 12\times 20/(12-1) \approx 21{,}818$.

Оценка генеральной дисперсии по второй выборке $\sigma_2^2 = n_2 s_2^2/(n_2-1) = 10\times 30/(10-1) \approx 33{,}333$.

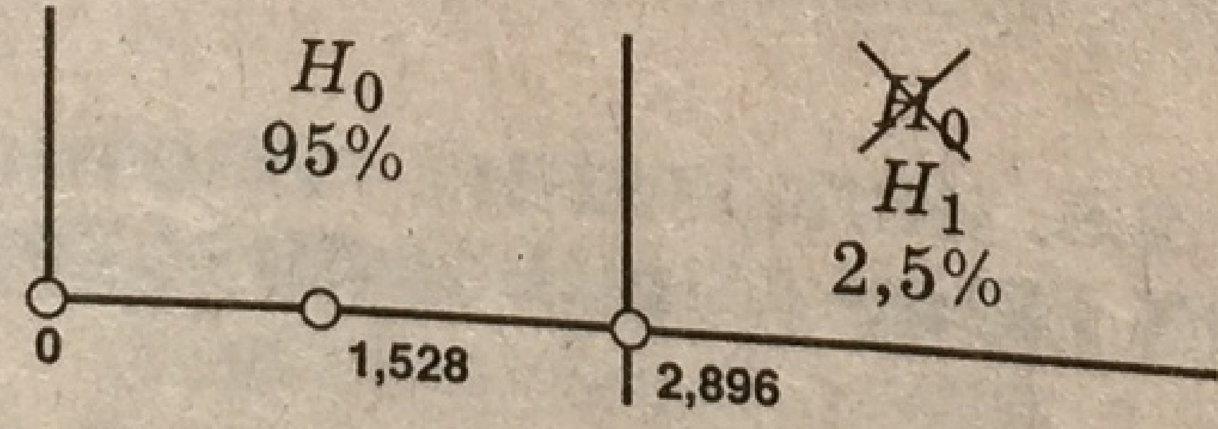
Статистика F = (бо́льшая оценка генеральной дисперсии)/(меньшая оценка генеральной дисперсии) = $33{,}333/21{,}818 \approx 1{,}528$.

Так как $33{,}333 > 21{,}818$, то $n_A = 10$, $n_B = 12$.

Проведем двустороннюю проверку.

$\alpha = (1-p)/2 = (1-0{,}95)/2 = 0{,}025 \Rightarrow F_{\alpha;n_A-1;n_B-1} = F_{0{,}025;10-1;12-1} = 2{,}896 \Rightarrow$ граничные точки $\pm 2{,}896$.

Отметим значения на числовой оси.



Принимаем гипотезу H_0 на уровне значимости 5%. Риски инвестиций равны.

Задача 11. Инвестиция 1 рассчитана на $n_1 = 14$ лет, дисперсия ежегодных прибылей $s_1^2 = (15\%)^2$. Инвестиция 2 рассчитана на $n_2 = 12$ лет, дисперсия ежегодных прибылей $s_2^2 = (20\%)^2$. Предполагается, что распределение ежегодных прибылей на инвестиции подчиняется нормальному

закону распределения. Равны ли риски инвестиций 1 и 2? Доверительная вероятность $p = 99\%$.

§ 2.5.1. Двухвыборочный F-тест для дисперсии

В Excel существует надстройка *Пакет анализа*, которая позволяет автоматически провести испытание гипотезы о двух генеральных дисперсиях. Нужно воспользоваться командой *Сервис*. В раскрывшемся списке команд должна присутствовать команда *Анализ данных*. При ее отсутствии необходимо выбрать команду *Надстройки* и поставить «галочку» рядом с командой *Пакет анализа*. Если же команда *Пакет анализа* отсутствует, то нужно произвести доустановку Excel.

Сервис → *Анализ данных* → *Двухвыборочный F-тест для дисперсии* → *OK*. Откроется диалоговое окно, которое нужно заполнить. В графе *Интервал переменной 1:* указывается ссылка на ячейки, содержащие значения первой выборки. В графе *Интервал переменной 2:* указывается ссылка на ячейки, содержащие значения второй выборки.

Если первая из ячеек содержит пояснительный текст, то рядом со словом *Метки* нужно поставить «галочку». В графе *Альфа* указывается уровень значимости $1 - p$ (по умолчанию там уже указано 0,05, но исследователь может выбрать и свое значение). Также указываются параметры вывода (*Выходной интервал*, *Новый рабочий лист*, *Новая рабочая книга*) → *OK*. Откроется итоговое окно.

Здесь проверка всегда односторонняя. Если в графе *P(F<=f) одностороннее* указана величина, меньшая выбранного *Альфа*, то мы отклоняем гипотезу H_0 на уровне значимости *Альфа*.

Двухвыборочный F-тест для дисперсии		
	Переменная 1	*Переменная 2*
Среднее	$\bar{X}_1$	$\bar{X}_2$
Дисперсия	σ_1^2	σ_2^2
Наблюдения	n_1	n_2
df	$n_1 - 1$	$n_2 - 1$
F	Статистика F	
P(F<=f) одностороннее		
F критическое одностороннее	$F_{\alpha; n_A - 1; n_B - 1}$	

Если же надстройки *Пакет анализа* нет, то можно воспользоваться статистической функцией ФТЕСТ (массив 1; массив 2) мастера функций f_x пакета Excel. Массив 1 и массив 2 — это ссылки на ячейки, содержащие значения двух выборок. Функция ФТЕСТ выдает значение доверительной вероятности p для принятия H_0 при двусторонней проверке. Затем исследователь решает, устраивает ли его такое значение доверительной вероятности.

§ 2.6. СРАВНЕНИЕ СРЕДНИХ ВЕЛИЧИН ДВУХ ВЫБОРОК ПРИ ИЗВЕСТНЫХ ГЕНЕРАЛЬНЫХ ДИСПЕРСИЯХ

Часто исследователя интересует, одинаковы или нет средние величины двух выборок, взятых из двух нормальных генеральных совокупностей. При этом известны генеральные дисперсии σ_1^2 и σ_2^2.

H_0: $a_1 = a_2$ (генеральные средние равны), то есть $a_1 - a_2 = 0$.

H_1: $a_1 \neq a_2$.

Граничные точки: z_α (для правосторонней проверки), $-z_\alpha$ (для левосторонней проверки), $\pm z_\alpha$ (для двусторонней проверки). Значение z_α находим по таблице (*см. § 1.1*).

Статистика $z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$, где $\overline{X}_1$ и $\overline{X}_2$ — выборочные средние этих выборок.

Пример 12. Автомат 1 и автомат 2 фасуют чай в пачки. Стандартные отклонения $\sigma_1 = 1$ г и $\sigma_2 = 2$ г соответственно. В случайной выборке объема $n_1 = 20$ пачек для автомата 1 средний вес $\overline{X}_1 = 101$ г. В случайной выборке объема $n_2 = 15$ пачек для автомата 2 средний вес $\overline{X}_2 = 98$ г. Верно ли, что оба автомата фасуют чай в пачки одинакового среднего веса? Доверительная вероятность $p = 95\%$.

H_0: $a_1 = a_2$ (средний вес одинаков), то есть $a_1 - a_2 = 0$.

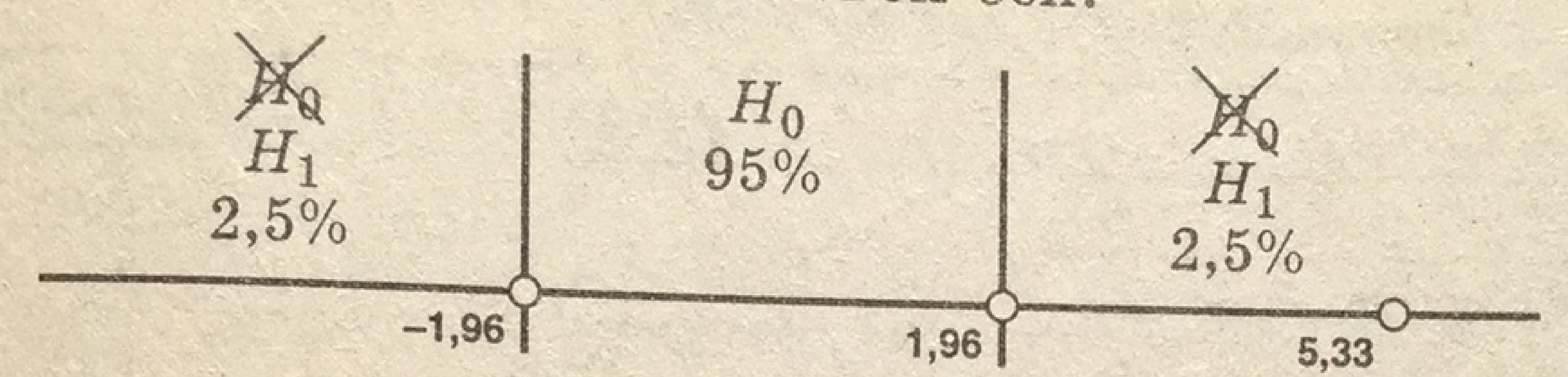
H_1: $a_1 \neq a_2$.

Проведем двустороннюю проверку.

$\alpha = (1-p)/2 = (1-0{,}95)/2 = 0{,}025 \Rightarrow z_\alpha = 1{,}96 \Rightarrow$ граничные точки 1,96.

Статистика $z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}} = \dfrac{101-98}{\sqrt{1^2/20 + 2^2/15}} \approx 5{,}33.$

Отметим значения на числовой оси.



Отклоняем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Средний вес пачек чая для этих автоматов различен.

Задача 12. Автомат 1 и автомат 2 фасуют чай в пачки. Стандартные отклонения $\sigma_1 = 0{,}5$ г и $\sigma_2 = 1$ гр. соответственно. В случайной выборке объема $n_1 = 12$ пачек для автомата 1 средний вес $\overline{X}_1 = 81$ г. В случайной выборке объема $n_2 = 16$ пачек для автомата 2 средний вес $\overline{X}_2 = 80$ гр. Верно ли, что оба автомата фасуют чай в пачки одинакового среднего веса? Доверительная вероятность $p = 99\%$.

§ 2.6.1. Двухвыборочный z-тест для средних (Excel)

Excel позволяет провести испытание гипотезы о равенстве средних двух нормальных распределений с известными генеральными дисперсиями.

Сервис → *Анализ данных* → *Двухвыборочный z-тест для средних* → *OK*. Раскроется диалоговое окно, которое нужно заполнить. В графе *Интервал переменной 1:* указывается ссылка на ячейки, содержащие значения первой выборки. В графе *Интервал переменной 2:* указывается ссылка на ячейки, содержащие значения второй выборки.

Если первая из ячеек содержит пояснительный текст, то рядом со словом *Метки* нужно поставить «галочку». В графе *Альфа* указывается уровень значимости $1-p$ (по умолчанию там уже указано 0,05, но исследователь может выбрать и свое значение). В графе *Гипотетическая средняя разность:* пишем 0. В графах *Дисперсия переменной 1 (известная):* и *Дисперсия переменной 2 (известная):* указываются значения σ_1^2 и σ_2^2 соответственно. Также указываются параметры вывода (*Выходной интервал*, *Новый рабочий лист*, *Новая рабочая книга*) → *OK*. Откроется итоговое окно.

В графах *P(Z<=z)* дано значение уровня значимости для односторонней и двусторонней проверок. Если это значение меньше заданного *Альфа*, то гипотеза H_0 отвергается.

Двухвыборочный z-тест для средних		
	Переменная 1	*Переменная 2*
Среднее	$\overline{X}_1$	$\overline{X}_2$
Известная дисперсия	σ_1^2	σ_2^2
Наблюдения	n_1	n_2
Гипотетическая разность средних	0	
Z	Статистика z	
P(Z<=z) односторонняя		
z критическое одностороннее	z_α для односторонней проверки	
P(Z<=z) двусторонняя		
z критическое двустороннее	z_α для двусторонней проверки	

§ 2.7. ИСПЫТАНИЕ ГИПОТЕЗЫ ПО ВЫБОРОЧНЫМ СРЕДНИМ ПРИ НЕИЗВЕСТНЫХ ГЕНЕРАЛЬНЫХ ДИСПЕРСИЯХ

Нужно определить, взяты ли две выборки объема n_1 и n_2 соответственно из нормальных генеральных совокупностей с одинаковыми средними.

H_0: $a_1 = a_2$ (генеральные средние равны), то есть $a_1 - a_2 = 0$.

$\overline{X}_1$ и $\overline{X}_2$ — выборочные средние для первой и второй выборок соответственно, s_1^2 и s_2^2 — выборочные дисперсии для первой и второй выборок соответственно. Вид граничных точек и статистики зависит от того, равны или нет между собой неизвестные генеральные дисперсии. Поэтому сначала надо проверить гипотезу о равенстве двух генеральных дисперсий (*см. § 2.5*).

§ 2.7.1. Случай равенства генеральных дисперсий

По таблице t-распределения находим $t_{\alpha;n_1+n_2-2}$. Граничные точки: $t_{\alpha;n_1+n_2-2}$ (для правосторонней проверки), $-t_{\alpha;n_1+n_2-2}$ (для левосторонней проверки), $\pm t_{\alpha;n_1+n_2-2}$ (для двусторонней проверки).

$$\text{Статистика } t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}.$$

Пример 13. Для производства каждой из $n_1 = 10$ деталей по первой технологии было затрачено в среднем $\overline{X}_1 = 30$ с (выборочная дисперсия $s_1^2 = 1$ с^2). Для производства каждой из $n_2 = 16$ деталей по второй технологии было затрачено в среднем $\overline{X}_2 = 28$ с (выборочная дисперсия $s_2^2 = 2$ с^2). Можно ли сделать вывод, что по первой технологии требуется в среднем больше времени для производства одной детали? Доверительная вероятность $p = 95\%$.

Применив результаты *§ 2.5*, получаем, что неизвестные генеральные дисперсии равны.

H_0: $a_1 = a_2$.

H_1: $a_1 > a_2$.

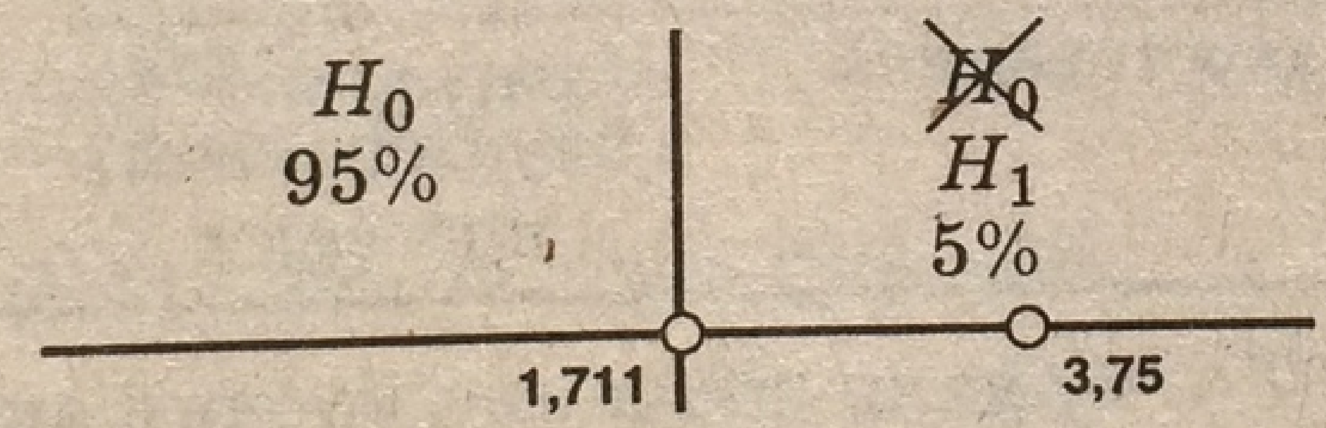
Проведем правостороннюю проверку.

$\alpha = 1 - p = 1 - 0{,}95 = 0{,}05 \Rightarrow t_{\alpha; n_1+n_2-2} = t_{0{,}05;10+16-2} = 1{,}711$. Это граничная точка.

$$\text{Статистика } t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\dfrac{n_1 s_1^2 + n_2 s_2^2}{n_1 + n_2 - 2}\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} =$$

$$= \frac{30 - 28}{\sqrt{\dfrac{10 \times 1 + 16 \times 2}{10 + 16 - 2}\left(\dfrac{1}{10} + \dfrac{1}{16}\right)}} \approx 3{,}75.$$

Отметим значения на числовой оси.



Отклоняем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. По первой технологии требуется в среднем больше времени для производства одной детали.

Задача 13. Для производства каждой из $n_1 = 12$ деталей по первой технологии было затрачено в среднем $\overline{X}_1 = 25$ с (выборочная дисперсия $s_1^2 = 1{,}5$ с^2). Для производства каждой из $n_2 = 11$ деталей по второй технологии было затрачено в среднем $\overline{X}_2 = 23$ с (выборочная дисперсия $s_2^2 = 2$ с^2). Можно ли сделать вывод, что по первой технологии требуется в среднем больше времени для производства одной детали? Доверительная вероятность $p = 99\%$.

§ 2.7.1.1. Двухвыборочный t-тест с одинаковыми дисперсиями

Excel позволяет провести испытание гипотезы о равенстве средних двух нормальных распределений в случае равенства неизвестных генеральных дисперсий.

Сервис → Анализ данных → Двухвыборочный t-тест с одинаковыми дисперсиями → OK. Раскроется диалоговое окно, которое нужно заполнить. *OK.* Откроется итоговое окно.

Двухвыборочный t-тест с одинаковыми дисперсиями		
	Переменная 1	*Переменная 2*
Среднее	$\overline{X}_1$	$\overline{X}_2$
Дисперсия	s_1^2	s_2^2
Наблюдения	n_1	n_2
Объединенная дисперсия		
Гипотетическая разность средних	0	
df	$n_1 + n_2 - 2$	
t-статистика	Статистика t	
P(T<=t) одностороння		
t критическое одностороннее	$t_{\alpha;n_1+n_2-2}$ для односторонней проверки	
P(T<=t) двусторонняя		
t критическое двустороннее	$t_{\alpha;n_1+n_2-2}$ для двусторонней проверки	

В графах *P(T<=t)* дано значение уровня значимости для односторонней и двусторонней проверок. Если это значение меньше заданного *Альфа*, то гипотеза H_0 отвергается.

Если надстройки *Пакет анализа* нет, то можно воспользоваться статистической функцией ТТЕСТ (массив 1; массив 2; хвосты; 2) мастера функций f_x пакета Excel. Для односторонней проверки хвосты = 1, для двусторонней проверки хвосты = 2. Функция ТТЕСТ выдает значение уровня значимости, которое нужно сравнить с соответствующим α.

§ 2.7.2. Случай неравенства генеральных дисперсий

В этом случае лучше обратиться к *Пакету анализа*. Но при $n_1 \geq 30$ и $n_2 \geq 30$ можно применить следующую схе-

му. Граничные точки: z_α (для правосторонней проверки), $-z_\alpha$ (для левосторонней проверки), $\pm z_\alpha$ (для двусторонней проверки). Значение z_α находим по таблице (*см. § 1.1*).

Статистика $z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1{}^2/(n_1-1) + s_2{}^2/(n_2-1)}}$, где $\overline{X}_1$ и $\overline{X}_2$ — выборочные средние этих выборок.

Пример 14. Для производства каждой из $n_1 = 51$ детали по первой технологии было затрачено в среднем $\overline{X}_1 = 30$ с (выборочная дисперсия $s_1{}^2 = 6$ с^2). Для производства каждой из $n_2 = 41$ детали по второй технологии было затрачено в среднем $\overline{X}_2 = 25$ с (выборочная дисперсия $s_2{}^2 = 3$ с^2). Можно ли сделать вывод, что по первой технологии требуется в среднем больше времени для производства одной детали? Доверительная вероятность $p = 95\%$.

Применив результаты *§ 2.5*, получаем, что неизвестные генеральные дисперсии различны.

H_0: $a_1 = a_2$.

H_1: $a_1 > a_2$.

Проведем правостороннюю проверку.

$\alpha = 1 - p = 1 - 0{,}95 = 0{,}05 \Rightarrow z_\alpha = 1{,}645$. Это граничная точка.

Статистика $z = \dfrac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_1{}^2/(n_1-1) + s_2{}^2/(n_2-1)}} =$

$= \dfrac{30 - 25}{\sqrt{6/(51-1) + 3/(41-1)}} \approx 11{,}323.$

Отметим значения на числовой оси.

H_0
95%
~~H_0~~
H_1
5%
1,645
11,323

Мы отклоняем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. По первой технологии требуется в среднем больше времени для производства одной детали.

Задача 14. Для производства каждой из $n_1 = 51$ детали по первой технологии было затрачено в среднем $\overline{X}_1 = 32$ с (выборочная дисперсия $s_1{}^2 = 9$ с^2). Для производства каждой из $n_2 = 41$ детали по второй технологии было затрачено в среднем $\overline{X}_2 = 28$ с (выборочная дисперсия $s_2{}^2 = 4$ с^2). Можно ли сделать вывод, что по первой технологии требуется в среднем больше времени для производства одной детали? Доверительная вероятность $p = 90\%$.

§ 2.7.2.1. Двухвыборочный t-тест с различными дисперсиями

Excel позволяет провести испытание гипотезы о равенстве средних двух нормальных распределений в случае неравенства неизвестных генеральных дисперсий.

Сервис → Анализ данных → Двухвыборочный t-тест различными с дисперсиями → ОК. Далее *см.* § *2.7.1.1*. Для этого случая ТТЕСТ (массив 1; массив 2; хвосты; 3).

§ 2.8. ИСПЫТАНИЕ ГИПОТЕЗЫ ПО ДВУМ ВЫБОРОЧНЫМ ДОЛЯМ

Нужно определить, взяты ли две выборки большого объема ($n_1 \geq 30$, $n_2 \geq 30$) с выборочными долями $\hat{p}_1$ и $\hat{p}_2$ из генеральных совокупностей с одинаковой генеральной долей.

H_0: $p_1 = p_2$ (генеральные доли одинаковы).

Граничные точки: z_α (для правосторонней проверки), $-z_\alpha$ (для левосторонней проверки), $\pm z_\alpha$ (для двусторонней проверки). Значение z_α находим по таблице (*см.* § *1.1*).

Статистика $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}}$, где $\overline{p}$ — выборочная доля для объединенной выборки.

Пример 15. Проводились испытания нового лекарства. В эксперименте участвовали $n_1 = 3000$ мужчин и $n_2 = 3500$ женщин. У 50 мужчин и 110 женщин наблюдались побочные эффекты. Можно ли утверждать, что побочные эффекты от нового лекарства у женщин возникают чаще, чем у мужчин? Доверительная вероятность $p = 95\%$.

Выборочные доли $\hat{p}_1 = 50/3000 \approx 0{,}017$, $\hat{p}_2 = 110/3500 \approx 0{,}031$.

H_0: $p_1 = p_2$ (генеральные доли одинаковы).

H_1: $p_1 < p_2$.

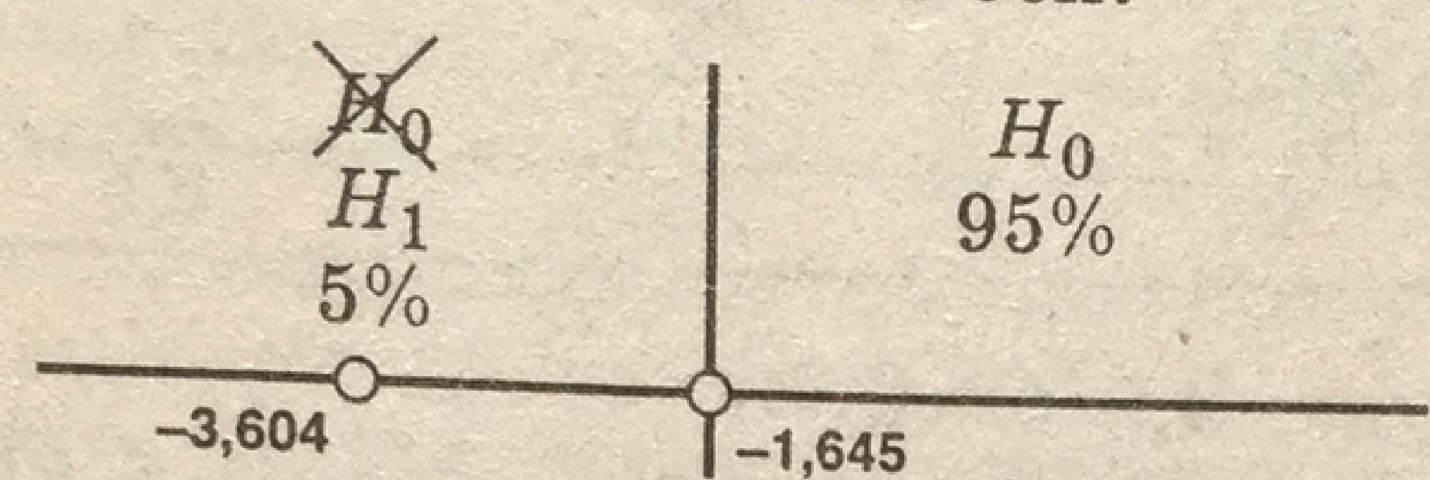
Проведем левостороннюю проверку.

$\alpha = 1 - p = 1 - 0{,}95 = 0{,}05 \Rightarrow z_\alpha = 1{,}645 \Rightarrow$ граничная точка $-1{,}645$. Выборочная доля для объединенной выборки равна $(50 + 110)/(3000 + 3500) \approx 0{,}025$.

Статистика $z = \dfrac{\hat{p}_1 - \hat{p}_2}{\sqrt{\overline{p}(1-\overline{p})\left(\dfrac{1}{n_1} + \dfrac{1}{n_2}\right)}} =$

$$= \frac{0{,}017 - 0{,}031}{\sqrt{0{,}025 \times (1 - 0{,}025)\left(\frac{1}{3000} + \frac{1}{3500}\right)}} \approx -3{,}604.$$

Отметим значения на числовой оси.



Отклоняем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Побочные эффекты от нового лекарства у женщин возникают чаще, чем у мужчин.

Задача 15. Проводились испытания нового лекарства. В эксперименте участвовали $n_1 = 2000$ мужчин и $n_2 = 2500$ женщин. У 40 мужчин и 70 женщин наблюдались побочные эффекты. Можно ли утверждать, что побочные эффекты от нового лекарства у женщин возникают чаще, чем у мужчин? Доверительная вероятность $p = 99\%$.

§ 2.9. ИСПЫТАНИЕ ГИПОТЕЗ ПО СПАРЕННЫМ ДАННЫМ

Иногда выборки не являются независимыми из-за наличия факторов, влияющих на выборки неизвестным путем. Тогда группируют элементы попарно (по одному из каждой выборки) и проводят испытание гипотезы на среднюю разностей между парными измерениями.

Пусть n — объем парной выборки. В каждой паре находим d — разность значений. Для полученных разностей ищем выборочную среднюю $\overline{X}_d$ и выборочное стандартное отклонение s_d. По таблице t-распределения находим $t_{\alpha;n-1}$.

Граничные точки: $t_{\alpha;n-1}$ (для правосторонней проверки), $-\,t_{\alpha;n-1}$ (для левосторонней проверки), $\pm t_{\alpha;n-1}$ (для двусторонней проверки).

Статистика $t = \dfrac{\overline{X}_d}{s_d/\sqrt{n-1}}$.

Пример 16. Можно ли утверждать, что шины заводов 1 и 2 имеют разную износоустойчивость? Доверительная вероятность $p = 95\%$.

Номер машины	X — расстояние для шин завода 1, тыс. км	Y — расстояние для шин завода 2, тыс. км	$d = X - Y$	d^2
1	60,2	59,4	0,8	0,64
2	62,3	58,3	4	16
3	61,3	62,1	−0,8	0,64
4	60,7	63,4	−2,7	7,29
5	63,4	60,8	2,6	6,76
Сумма	—	—	3,9	31,33

$\overline{X}_d = (\sum d)/n = 3{,}9/5 = 0{,}78.$

$s_d^2 = (\sum d^2)/n - (\overline{X}_d)^2 = 31{,}33/5 - 0{,}78^2 = 5{,}6576.$

$s_d \approx 2{,}379.$

H_0: средняя $a_d = 0$ (нет разницы между шинами).

H_1: $a_d \neq 0$ (есть разница)

Проведем двустороннюю проверку.

$\alpha = (1-p)/2 = (1-0{,}95)/2 = 0{,}025 \Rightarrow t_{\alpha;n-1} = t_{0{,}025;5-1} = 2{,}776 \Rightarrow$ граничные точки $\pm 2{,}776$.

$$\text{Статистика } t = \frac{\overline{X}_d}{s_d/\sqrt{n-1}} = \frac{0{,}78}{2{,}379/\sqrt{5-1}} \approx 1{,}311.$$

Отметим значения на числовой оси.

~~H_0~~ H_1 2,5% | H_0 95% | ~~H_0~~ H_1 2,5%

−2,776 | 1,311 | 2,776

Мы принимаем гипотезу H_0 на уровне значимости 5%. Износоустойчивость шин одинакова.

Задача 16. Можно ли утверждать, что шины заводов 1 и 2 имеют разную износоустойчивость? Доверительная вероятность $p = 99\%$.

Номер машины	X — расстояние для шин завода 1, тыс. км	Y — расстояние для шин завода 2, тыс. км
1	62,4	61,8
2	61,8	62,3
3	63,2	60,6
4	57,4	59,2
5	59,6	62,1

§ 2.9.1. Парный двухвыборочный t-тест для средних

Excel позволяет провести парный тест. *Сервис → Анализ данных → Парный двухвыборочный t-тест для средних → OK*. Раскроется диалоговое окно, которое нужно заполнить. *OK*. Появляется итоговое окно.

Парный двухвыборочный t-тест для средних		
	Переменная 1	*Переменная 2*
Среднее	$\bar{X}_1$	$\bar{X}_2$
Дисперсия		
Наблюдения	n	n
Корреляция Пирсона		
Гипотетическая разность средних	a_d	
df	$n - 1$	
t-статистика	Статистика t	
P(T<=t) односторонняя		
t критическое одностороннее	$t_{\alpha;n-1}$ для односторонней проверки	
P(T<=t) двусторонняя		
t критическое двустороннее	$t_{\alpha;n-1}$ для двусторонней проверки	

В графах *P(T<=t)* дано значение уровня значимости для односторонней и двусторонней проверок. Если это значение меньше заданного *Альфа*, то гипотеза H_0 отвергается.

Если же надстройки *Пакет анализа* нет, то можно воспользоваться статистической функцией ТТЕСТ (массив 1; массив 2; хвосты; 1) мастера функций f_x пакета Excel.

§ 2.10. НЕПАРАМЕТРИЧЕСКИЕ ИСПЫТАНИЯ

До сих пор мы предполагали, что генеральные совокупности распределены нормально или приблизительно нормально. Теперь мы откажемся от этих условий. Будем проверять гипотезу о наличии связи между значениями двух величин.

H_0: нет связи между значениями двух величин.

H_1: есть связь между значениями двух величин.

Составляется *таблица наблюдаемых частот*. По строкам изменяются значения первой величины, по столбцам — значения второй величины. В клетке с индексами (i, j) записана частота $f_0 = n_{ij}$ — число элементов, у которых значения первой и второй величин равны i и j соответственно. f_0 — наблюдаемая частота события.

По таблице наблюдаемых частот строят показанным далее способом *таблицу ожидаемых частот*. f_E — ожидаемая частота события. Должно выполняться условие $f_E \geq 5$ для каждой клетки таблицы, иначе надо объединить какие-то строки или столбцы.

Таблицу наблюдаемых частот и таблицу ожидаемых частот часто называют *таблицами сопряженности*. Пусть n — общее число наблюдений. m = (число строк таблицы − 1) × × (число столбцов таблицы − 1). Если таблица сопряженности содержит только одну строку, то m = число столбцов таблицы − 1.

Доверительная вероятность p, уровень значимости $\alpha = 1 - p$. Для α и m по таблице χ^2-распределения находим $\chi^2_{\alpha,m}$. Это граничная точка.

Статистика $\chi^2 = \sum \frac{(f_0 - f_E)^2}{f_E}$.

Если таблица сопряженности имеет размер 2×2, то вводится поправка Йетса: $\chi^2 = \sum \frac{(|f_0 - f_E| - 0,5)^2}{f_E}$.

Отметим значения на числовой оси. Для нахождения $\chi^2_{\alpha,m}$ можно воспользоваться статистической функцией ХИ2ОБР(α; m) мастера функций f_x пакета Excel.

Пример 17. Студенты сдавали экзамены по математике и физике. Есть ли связь между результатами экзаменов?

Результаты экзамена по математике	Результаты экзамена по физике			
	пять	четыре	три	два
пять	25	18	10	5
четыре	20	16	15	6
три	15	20	22	13
два	8	10	7	15

Приведенная таблица — это таблица наблюдаемых частот f_0. В клетке (1, 1) написано число 25, то есть 25 студентов получили и по физике, и по математике отличные оценки. В клетке (4, 2) написано число 10, то есть 10 студентов получили хорошие оценки по физике и неудовлетворительные оценки по математике. И т. д.

H_0: нет связи между оценками по математике и физике.

H_1: есть связь между оценками по математике и физике.

Построим таблицу ожидаемых частот f_E. Для этого суммируем числа по строкам и столбцам.

Результаты экзамена по математике	Результаты экзамена по физике				Сумма
	пять	четыре	три	два	
пять	25	18	10	5	58
четыре	20	16	15	6	57
три	15	20	22	13	70
два	8	10	7	15	40
Сумма	68	64	54	39	225

Всего получены результаты экзаменов $n = 225$ студентов. Отличный результат по математике показали 58 студентов, то есть доля тех, кто получил отличные оценки по математике, равна 58/225. Если верна гипотеза H_0, то можно ожидать, что 58/225 из 68 студентов, получивших по физике отличные оценки, показали отличные знания и по математике.

Аналогично можно рассчитать и другие ожидаемые частоты.

Результаты экзамена по математике	Результаты экзамена по физике				Сумма
	пять	четыре	три	два	
пять	68×58/225	64×58/225	54×58/225	39×58/225	58
четыре	68×57/225	64×57/225	54×57/225	39×57/225	57
три	68×70/225	64×70/225	54×70/225	39×70/225	70
два	68×40/225	64×40/225	54×40/225	39×40/225	40
Сумма	68	64	54	39	225

Ожидаемые частоты нельзя округлять до целого значения.

Результаты экзамена по математике	Результаты экзамена по физике				Сумма
	пять	четыре	три	два	
пять	17,5	16,5	13,9	10,1	58
четыре	17,2	16,2	13,7	9,9	57
три	21,2	19,9	16,8	12,1	70
два	12,1	11,4	9,6	6,9	40
Сумма	68	64	54	39	225

Если в какой-то клетке получилось значение, меньшее 5, то с целью уничтожения этого в таблицах нужно объединить какие-то строки или столбцы. При округлении надо следить, чтобы исходные суммы не изменились.

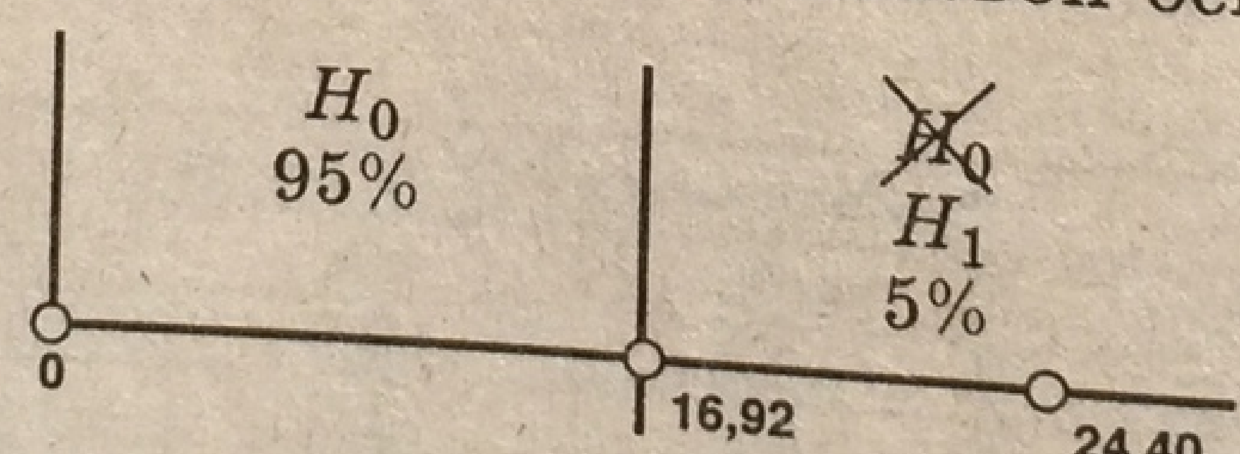
Доверительная вероятность $p = 0{,}95$, уровень значимости $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$. $m =$ (число строк таблицы $- 1$) $\times$ $\times$ (число столбцов таблицы $- 1$) $= (4 - 1) \times (4 - 1) = 9$.

Для α и m по таблице χ^2-распределения находим $\chi^2_{\alpha,m} = \chi^2_{0{,}05;9} = 16{,}92$. Это граничная точка. Найдем значение статистики χ^2.

f_0	f_E	$f_0 - f_E$	$(f_0 - f_E)^2$	$(f_0 - f_E)^2/f_E$
25	17,5	7,5	56,25	3,21
20	17,2	2,8	7,84	0,46
15	21,2	−6,2	38,44	1,81
8	12,1	−4,1	16,81	1,39
18	16,5	1,5	2,25	0,14
16	16,2	−0,2	0,04	0,00
20	19,9	0,1	0,01	0,00
10	11,4	−1,4	1,96	0,17
10	13,9	−3,9	15,21	1,09
15	13,7	1,3	1,69	0,12
22	16,8	5,2	27,04	1,61
7	9,6	−2,6	6,76	0,70
5	10,1	−5,1	26,01	2,58
6	9,9	−3,9	15,21	1,54
13	12,1	0,9	0,81	0,07
15	6,9	8,1	65,61	9,51
Сумма	—	0	—	24,40

Поясним, как заполняется таблица. Наблюдаемые частоты f_0 пишем в 1-м столбце, а соответствующие им ожидаемые частоты f_E – во 2-м столбце. Далее производим над столбцами действия, указанные в 1-й строке.

Статистика $\chi^2 = \sum \frac{(f_0 - f_E)^2}{f_E} = 24{,}40$ (сумма чисел 5-го столбца). Отметим значения на числовой оси.



Мы отклоняем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Есть связь между оценками, полученными студентами на экзаменах по математике и физике.

Замечание. Вместо заполнения последней таблицы можно воспользоваться статистической функцией ХИ2ТЕСТ мастера функций f_x пакета Excel. f_x → *статистические* → *ХИ2ТЕСТ*. Появляется диалоговое окно. В графе *фак-*

тический интервал указывается ссылка на ячейки, в которых хранятся наблюдаемые частоты. В графе *ожидаемый интервал* указывается ссылка на ячейки, в которых хранятся ожидаемые частоты. *OK*. Если полученное значение превышает уровень значимости $\alpha = 1 - p$, то гипотеза H_0 отклоняется.

Задача 17. Студенты сдавали экзамены по математике и физике. Есть ли связь между результатами экзаменов? Доверительная вероятность 99%.

Результаты экзамена по математике	Результаты экзамена по физике			
	пять	четыре	три	два
пять	20	17	12	6
четыре	22	15	17	5
три	21	19	20	12
два	9	8	7	18

Глава 3

ПАРНАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Очень часто исследователя интересует связь между переменными. Это помогает при анализе их поведения. Начнем с двух переменных, а в следующей главе рассмотрим случай многих переменных. Будет разработана модель для описания связи между переменными с математической точки зрения. Начнем с наиболее простых для анализа линейных уравнений.

§ 3.1. ПРОСТАЯ МОДЕЛЬ ЛИНЕЙНОЙ РЕГРЕССИИ

Существует или нет линейная связь между двумя переменными x, y? Проводим случайную выборку. При значениях $x_1, x_2, \ldots, x_n$ мы наблюдаем значения $y_1, y_2, \ldots, y_n$ соответственно. На плоскости Oxy отметим точки с координатами $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$.

Предположим, что точки группируются вокруг некоторой прямой линии $y = a + bx$. Тогда:

$$b = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}, \quad a = \frac{\sum_{i=1}^{n} y_i - b\sum_{i=1}^{n} x_i}{n}.$$

Точки не находятся точно на линии $y = a + bx$. Но это неудивительно. Ведь помимо x на поведение y оказывают влияние и другие факторы. Дальнейший анализ полученного уравнения позволяет сказать, насколько сильно влияние неучтенных факторов, действительно ли модель линейна и т. д. На переменные x, y накладывается ряд условий. Для описания природы связи используется термин *«регрессия»*. Коэффициент b называется *показателем наклона линии линейной регрессии*.

Пример 18. Изучается зависимость себестоимости единицы изделия (y, тыс. руб.) от величины выпуска продукции (x, тыс. шт.) по группам предприятий за отчетный период. Экономист обследовал $n = 5$ предприятий и получил следующие результаты (1-й и 2-й столбцы). Полагая, что между переменными x, y имеет место линейная зависимость, определим выборочное уравнение линейной регрессии.

Заполним таблицу.

Номер	x	y	x^2	xy
1	2	1,9	4	3,8
2	3	1,7	9	5,1
3	4	1,8	16	7,2
4	5	1,6	25	8
5	6	1,4	36	8,4
Сумма	20	8,4	90	32,5

$$b = \frac{n\sum\limits_{i=1}^{n} x_i y_i - \sum\limits_{i=1}^{n} x_i \sum\limits_{i=1}^{n} y_i}{n\sum\limits_{i=1}^{n} x_i^2 - \left(\sum\limits_{i=1}^{n} x_i\right)^2} = \frac{5 \times 32{,}5 - 20 \times 8{,}4}{5 \times 90 - 20^2} = -0{,}11.$$

$$a = \frac{\sum\limits_{i=1}^{n} y_i - b\sum\limits_{i=1}^{n} x_i}{n} = \frac{8{,}4 - (-0{,}11) \times 20}{5} = 2{,}12.$$

$$y = a + bx = 2{,}12 + (-0{,}11)x.$$

Задача 18. Фирма провела рекламную кампанию. Через 10 недель фирма решила проанализировать эффективность этого вида рекламы, сопоставив недельные объемы продаж (y, тыс. руб.) с расходами на рекламу (x, тыс. руб.).

x	5	8	6	5	3	9	12	4	3	10
y	72	76	78	70	68	80	82	65	62	90

Полагая, что между переменными x, y имеет место линейная зависимость, определить выборочное уравнение линейной регрессии.

Замечание. Вместо вычислений коэффициентов a и b по формулам можно воспользоваться соответственно статистическими функциями ОТРЕЗОК (изв_знач_y; изв_знач_x) и НАКЛОН (изв_знач_y; изв_знач_x) мастера функций f_x пакета Excel. Здесь изв_знач_y и изв_знач_x — это ссылки на ячейки, содержащие значения переменных y и x соответственно.

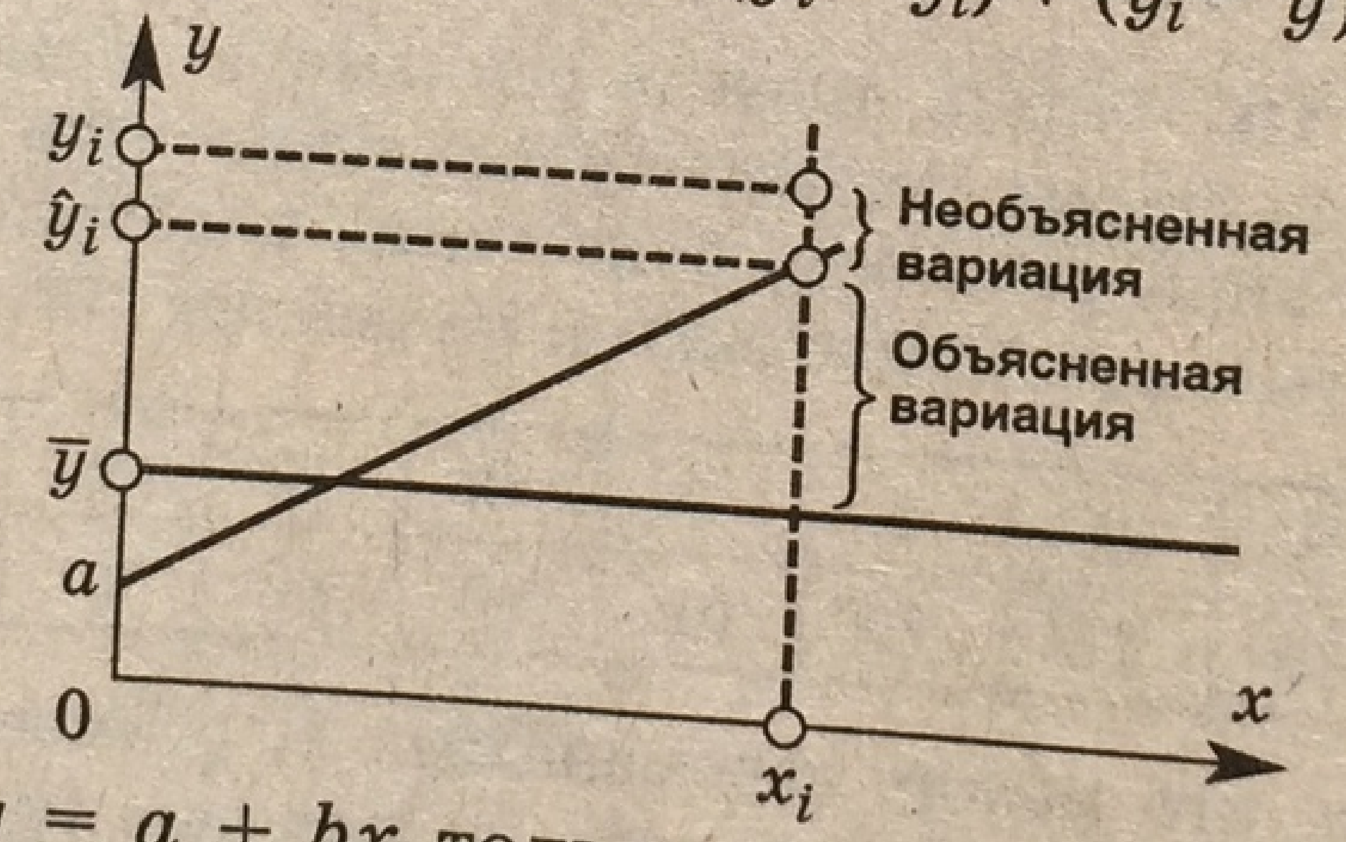
Обозначим через $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$ и $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ средние значения переменных y и x соответственно.

§ 3.2. ОШИБКИ

Проводим случайную выборку. При значениях $x_1, x_2, \ldots, x_n$ мы наблюдаем значения $y_1, y_2, \ldots, y_n$ соответственно. Получено уравнение $\hat{y} = a + bx$. Если вместо x подставить в это уравнение значения $x_1, x_2, \ldots, x_n$, то будут получены значения $\hat{y}_1, \ldots, \hat{y}_n$, которые, вообще говоря, будут отличаться от $y_1, y_2, \ldots, y_n$. Разница $y_i - \hat{y}_i = e_i$ называется *ошибкой* (остатком, отклонением). Значения коэффициентов a и b в уравнении $y = a + bx$, которые рассчитывались по приведенным в *§ 3.1* формулам, подбирались так, чтобы минимизировать сумму $\sum_{i=1}^{n} e_i^2$. Говорят, что они получены *методом наименьших квадратов* (МНК).

§ 3.3. КОЭФФИЦИЕНТ КОРРЕЛЯЦИИ ПИРСОНА. КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

Мы хотим знать, насколько хорошо приближает наши данные линейная модель. $y_i - \overline{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \overline{y}) = (\hat{y}_i - \overline{y}) + e_i$.



Формула $y = a + bx$ только частично объясняет вариацию значений y (а именно, слагаемое $\hat{y}_i - \overline{y}$). Но ведь на y влияют и другие факторы. Их влияние скрыто в остатке e_i. Если бы связь была строго линейной, то $e_i = 0$. И так для каждой точки x_i. $\sum_{i=1}^{n}(y_i - \overline{y})^2$ — это общая вариация переменной y. $\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$ — это вариация переменной y, которая объясняется формулой $y = a + bx$. $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ — это ва-

риация переменной y, которая не объясняется формулой $y = a + bx$.

Введем характеристику $r^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$ — *коэффициент детерминации*. Эта мера показывает величину вариации переменной y, которая объясняется переменной x при наличии линейной связи этих величин. В случае строгой линейной зависимости между x и y $r^2 = 1$. Если зависимость между x и y отсутствует, то $r^2 = 0$.

Коэффициент корреляции Пирсона:

$$r = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{\left(n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right)\left(n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2\right)}}. \ |r| \le 1.$$

Вторая дробь — удобная расчетная формула, которую чаще всего используют.

Коэффициент корреляции Пирсона r содержит информацию о поведении y с ростом x. Знак коэффициента корреляции Пирсона r совпадает со знаком коэффициента b. Чем ближе r к 1, тем ближе связь между x и y к линейной. При $r = 0$ линейной связи между x и y не существует (но, возможно, между x и y есть другая зависимость).

Пример 19. Найдем остатки e_i, коэффициент корреляции Пирсона и коэффициент детерминации в примере 18. $y = 2{,}12 - 0{,}11x$.

Номер	x	y	y^2	$\hat{y} = 2{,}12 - 0{,}11x$	$e = y - \hat{y}$
1	2	1,9	3,61	1,90	0,00
2	3	1,7	2,89	1,79	–0,09
3	4	1,8	3,24	1,68	0,12
4	5	1,6	2,56	1,57	0,03
5	6	1,4	1,96	1,46	–0,06
Сумма	20	8,4	14,26		

$$r = \frac{n\sum_{i=1}^{n}x_i y_i - \sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{\sqrt{\left(n\sum_{i=1}^{n}x_i^2 - \left(\sum_{i=1}^{n}x_i\right)^2\right)\left(n\sum_{i=1}^{n}y_i^2 - \left(\sum_{i=1}^{n}y_i\right)^2\right)}} =$$

$$= \frac{5 \times 32{,}5 - 20 \times 8{,}4}{\sqrt{(5 \times 90 - 20^2)(5 \times 14{,}26 - 8{,}4^2)}} \approx -0{,}904.$$

Это значение близко к -1, что свидетельствует об очень сильной отрицательной связи (с ростом x значения y убывают). Знаки $b = -0{,}11$ и $r = -0{,}904$ совпадают. Коэффициент детерминации $r^2 = (-0{,}904)^2 \approx 0{,}817$, то есть 81,7% общей вариации себестоимости y зависит от выпуска продукции x. Наша модель не объясняет 18,3% вариации себестоимости. Эта часть вариации объясняется факторами, не включенными в модель.

Задача 19. Найти остатки e_i, коэффициент корреляции Пирсона и коэффициент детерминации в задаче 18.

Замечание. Для вычисления коэффициента корреляции Пирсона можно воспользоваться статистическими функциями ПИРСОН (массив 1; массив 2) или КОРРЕЛ (массив 1; массив 2) мастера функций f_x пакета Excel. Массив 1 и массив 2 — это ссылки на ячейки, содержащие значения переменных. Для вычисления коэффициента детерминации можно воспользоваться статистической функцией КВПИРСОН (изв_знач_y; изв_знач_x).

§ 3.4. ПРЕДСКАЗАНИЯ И ПРОГНОЗЫ НА ОСНОВЕ ЛИНЕЙНОЙ МОДЕЛИ РЕГРЕССИИ

Мы можем воспользоваться построенной моделью для нахождения значения y при известном значении x. Модель строилась по значениям $x_1, x_2, \ldots, x_n$. Поэтому поиск значения y для x из интервала (x_1, x_n) называется *предсказанием*, а поиск значения y для x вне интервала (x_1, x_n) называется *прогнозом*. Чем дальше расположен x от интервала (x_1, x_n), тем менее точным будет прогноз.

Пример 20. Найдем ожидаемое значение себестоимости y при выпуске продукции $x = 5{,}5$ тыс. шт. $y = 2{,}12 - 0{,}11x$.

Тогда $y(5{,}5) = 2{,}12 - 0{,}11 \times 5{,}5 = 1{,}515$ тыс. руб.

Задача 20. Найти ожидаемое значение еженедельного объема продаж y при расходах на рекламу $x = 5{,}5$ тыс. руб. в задаче 18.

Замечание. Для прогноза значений переменной y можно воспользоваться статистической функцией ТЕНДЕНЦИЯ (изв_знач_y; изв_знач_x; нов_знач_x; константа) мастера функций f_x пакета Excel. Нов_знач_x — это ссылка на ячейки, содержащие значения переменной x, для кото-

рых ищется прогноз. Если необязательный аргумент константа = 0, то коэффициент $a = 0$. По известным значениям переменных x, y функция сама подбирает уравнение прямой линии и дает прогноз. Функцию ТЕНДЕНЦИЯ можно использовать и в случае множественной линейной регрессии. Для парной линейной регрессии можно воспользоваться и статистической функцией ПРЕДСКАЗ (x; изв_знач_y; изв_знач_x), где x — это значение переменной x, для которого ищется прогноз.

§ 3.5. ОСНОВНЫЕ ПРЕДПОСЫЛКИ МОДЕЛИ ПАРНОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Основные предпосылки:

1) связь между переменными x, y является линейной;

2) независимая переменная x может быть использована для прогноза y;

3) остатки (то есть ошибки) нормально распределены;

4) для всех данных x математическое ожидание ошибки равно нулю и дисперсия ошибки постоянна;

5) ошибки независимы.

§ 3.6. ИСПЫТАНИЕ ГИПОТЕЗЫ ДЛЯ ОЦЕНКИ ЛИНЕЙНОСТИ СВЯЗИ

Между переменными x, y предполагается наличие линейной связи $y = \alpha + \beta x + \varepsilon$, где ошибка ε — это отклонение значения y от линии $y = \alpha + \beta x$. Мы производим парную выборку значений переменных x, y и методом наименьших квадратов получаем оценки коэффициентов α и β — a и b соответственно. Очевидно, что для другой выборки оценки a и b будут другими. Как, зная оценки a и b, убедиться, что связь между переменными x, y действительно линейная?

§ 3.6.1. Испытание гипотезы для оценки линейности связи на основе оценки коэффициента корреляции в генеральной совокупности

Показатель наличия линейной связи в генеральной совокупности — это коэффициент корреляции. Для генеральной совокупности он равен ρ. Нам это значение неизвестно. По данным выборки мы получаем оценку для ρ — вы-

борочный коэффициент корреляции r — и на основании r проводим испытание гипотезы о наличии линейной связи между переменными x, y в генеральной совокупности. Наш вывод о наличии линейной связи между переменными x, y в генеральной совокупности зависит от объема выборки. Чем больше объем нашей выборки, тем надежнее полученный результат.

H_0: $\rho = 0$, то есть между переменными x, y отсутствует линейная связь в генеральной совокупности.

H_1: $\rho \neq 0$, то есть между переменными x, y есть линейная связь в генеральной совокупности.

Задается доверительная вероятность p. Пусть n — объем парной выборки. Двусторонняя проверка. $\alpha = (1 - p)/2$.

По таблице t-распределения находим $t_{\alpha;n-2}$. Граничные точки $\pm t_{\alpha;n-2}$.

Статистика $t = \sqrt{r^2(n-2)/(1-r^2)}$.

Пример 21. Вернемся к примерам 18, 19. Проверим гипотезу о наличии линейной связи между переменными x, y в генеральной совокупности. Доверительная вероятность $p = 95\%$. $n = 5$.

H_0: $\rho = 0$, то есть между переменными x, y отсутствует линейная связь в генеральной совокупности.

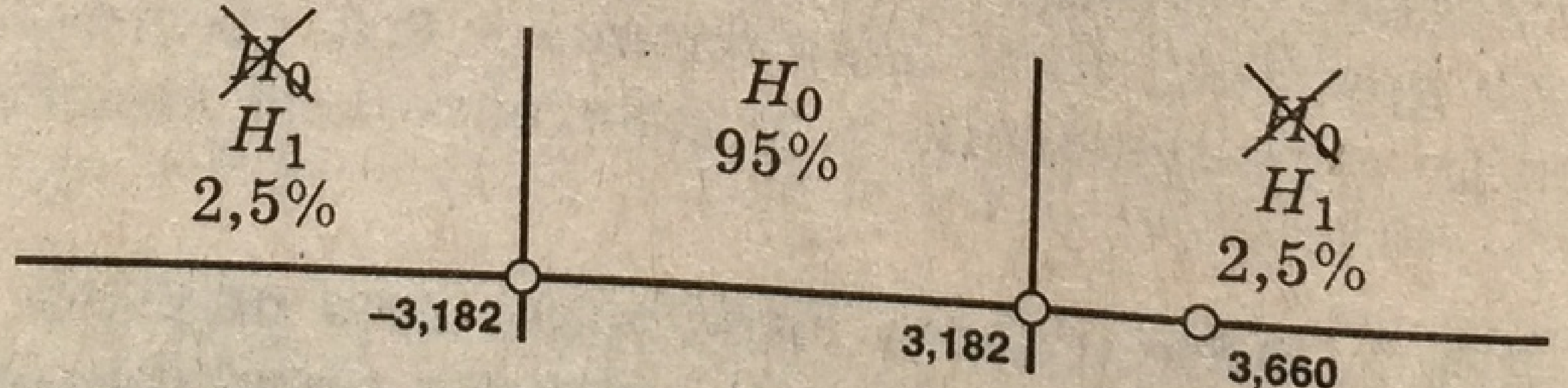
H_1: $\rho \neq 0$, то есть между переменными x, y есть линейная связь в генеральной совокупности.

Проведем двустороннюю проверку.

$\alpha = (1 - p)/2 = (1 - 0{,}95)/2 = 0{,}025$. По таблице t-распределения находим $t_{\alpha;n-2} = t_{0,025;5-2} = 3{,}182$. Граничные точки $\pm 3{,}182$.

Статистика $t = \sqrt{r^2(n-2)/(1-r^2)} = \sqrt{0{,}817 \times (5-2)/(1-0{,}817)} \approx 3{,}660$.

Отметим значения на числовой оси.



Мы отклоняем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Между переменными x, y есть линейная связь в генеральной совокупности.

Задача 21. В задачах 18, 19 проверить гипотезу о наличии линейной связи между переменными x, y в генеральной совокупности. Доверительная вероятность $p = 99\%$.

§ 3.6.2. Испытание гипотезы для оценки линейности связи на основе показателя наклона линейной регрессии

В случае парной линейной регрессии функция показателя наклона β аналогична функции коэффициента корреляции. Поэтому нужно ограничиться только одной проверкой.

H_0: $\beta = 0$, то есть между переменными x, y отсутствует линейная связь в генеральной совокупности.

H_1: $\beta \neq 0$, то есть между переменными x, y есть линейная связь в генеральной совокупности.

Задается доверительная вероятность p. n — объем парной выборки. Двусторонняя проверка. $\alpha = (1-p)/2$.

По таблице t-распределения находим $t_{\alpha;n-2}$. Граничные точки $\pm t_{\alpha;n-2}$.

Дисперсия распределения остатков вдоль линии регрессии $S^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2}$, S — стандартная ошибка.

Стандартная ошибка коэффициента b:

$$S_b = \frac{S}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}} = \frac{S}{\sqrt{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}}.$$

Статистика $t = b/S_b$.

Пример 22. Вернемся к примерам 18, 19. Проверим гипотезу о наличии линейной связи между переменными x, y в генеральной совокупности. Доверительная вероятность $p = 95\%$. $n = 5$.

H_0: $\beta = 0$, то есть между переменными x, y отсутствует линейная связь в генеральной совокупности.

H_1: $\beta \neq 0$, то есть между переменными x, y есть линейная связь в генеральной совокупности.

Проведем двустороннюю проверку.

$\alpha = (1-p)/2 = (1-0{,}95)/2 = 0{,}025$.

По таблице t-распределения находим $t_{\alpha;n-2} = t_{0,025;5-2} = 3{,}182$. Граничные точки $\pm 3{,}182$.

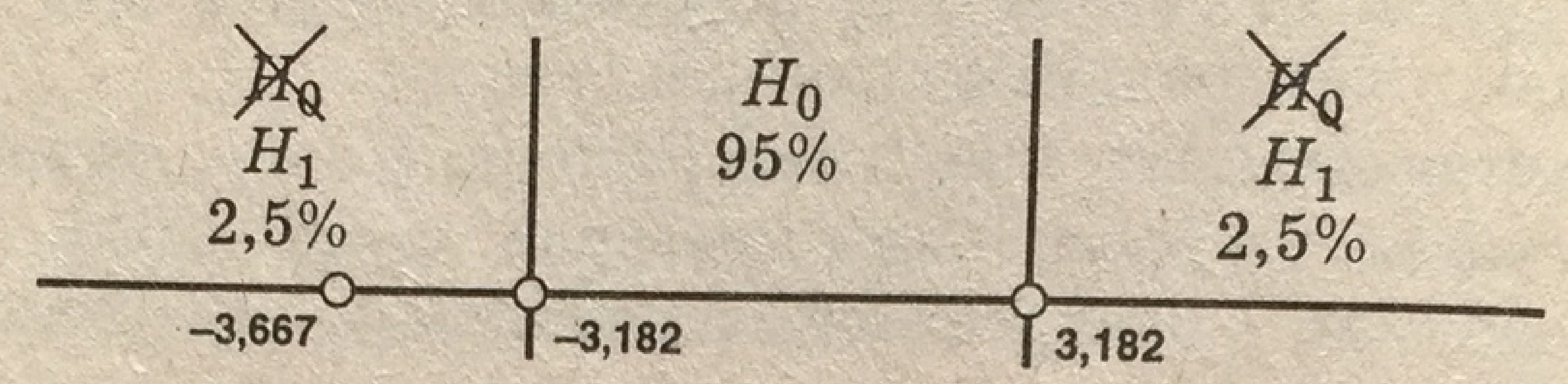
Номер	e_i	e_i^2
1	0	0
2	–0,09	0,0081
3	0,12	0,0144
4	0,03	0,0009
5	–0,06	0,0036
Сумма		0,0270

$$S^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-2} = \frac{0{,}027}{5-2} = 0{,}009. \quad S \approx 0{,}095.$$

$$S_b = \frac{S}{\sqrt{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}} \approx \frac{0{,}095}{\sqrt{90 - 20^2/5}} \approx 0{,}03.$$

Статистика $t = b/\ S_b = -0{,}11/0{,}03 \approx -3{,}667$.

Отметим значения на числовой оси.



Мы отклоняем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Между переменными x, y есть линейная связь в генеральной совокупности.

Задача 22. В задачах 18,19 проверить гипотезу о наличии линейной связи между переменными x,y в генеральной совокупности на основе показателя наклона. Доверительная вероятность $p = 99\%$.

Замечание. Для расчета стандартной ошибки вместо формулы $S = \sqrt{\frac{\sum_{i=1}^{n} e_i^2}{n-2}}$ можно воспользоваться статистической функцией СТОШYX (изв_знач_y; изв_знач_x) мастера функций f_x пакета Excel.

§ 3.7. ДОВЕРИТЕЛЬНЫЕ ИНТЕРВАЛЫ В ЛИНЕЙНОМ РЕГРЕССИОННОМ АНАЛИЗЕ

Проведя испытания гипотез (*§ 3.6*), мы пришли к выводу, что связь между переменными x, y линейна и задается неизвестной нам формулой $y = \alpha + \beta x$. Мы производим парную выборку значений переменных x, y и методом наименьших квадратов получаем оценки коэффициентов α и β — a и b соответственно. Получена формула $y = a + bx$, которой мы можем воспользоваться для оценки значений y при заданном значении x. По полученным точечным оценкам строят доверительные интервалы. Обычно это доверительные интервалы для показателя наклона линии линей-

ной регрессии β, для среднего значения y при заданном значении x и для значений y при заданном значении x. Задается доверительная вероятность p. n — объем парной выборки. $\alpha = (1-p)/2$. По таблице t-распределения находим $t_{\alpha;n-2}$.

§ 3.7.1. Доверительный интервал для показателя наклона линии линейной регрессии

Доверительный интервал имеет вид $b \pm t_{\alpha;n-2}S_b$, где S_b — стандартная ошибка коэффициента b.

Пример 23. Вернемся к примерам 18 и 22. Найдем доверительный интервал для показателя наклона линии линейной регрессии. Доверительная вероятность $p = 95\%$.

$b \pm t_{\alpha;n-2}S_b = -0{,}11 \pm 3{,}182 \times 0{,}03 \approx -0{,}11 \pm 0{,}10$, то есть $-0{,}21 < \beta < -0{,}01$.

Задача 23. В задачах 18 и 22 найти доверительный интервал для показателя наклона линии линейной регрессии. Доверительная вероятность $p = 99\%$.

§ 3.7.2. Доверительный интервал для среднего значения переменной *y* при данном значении переменной *x*

Обозначим данное значение переменной x через x_0. Доверительный интервал для среднего значения переменной y при x_0 задается формулой:

$$y = a + bx_0 \pm t_{\alpha;n-2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}},$$

где S — стандартная ошибка.

Чем больше величина $|x_0 - \overline{x}|$, тем шире доверительный интервал.

Пример 24. Вернемся к примерам 18 и 22. Найдем доверительный интервал для среднего значения переменной y при заданном значении $x_0 = 5{,}5$ тыс. шт. Доверительная вероятность $p = 95\%$.

$$\overline{x} = \sum_{i=1}^{n} x_i/n = 20/5 = 4.$$

$$y = a + bx_0 \pm t_{\alpha;n-2}S\sqrt{\frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n}} =$$

$$= 2{,}12 - 0{,}11 \times 5{,}5 \pm 3{,}182 \times 0{,}095 \sqrt{\frac{1}{5} + \frac{(5{,}5-4)^2}{90 - 20^2/5}} \approx$$

$$\approx 1{,}515 \pm 0{,}197.$$

То есть доверительный интервал для среднего значения переменной y при заданном значении $x_0 = 5{,}5$ тыс. шт. равен (1,318; 1,712).

Задача 24. В задачах 18 и 22 найти доверительный интервал для среднего значения переменной y при заданном значении $x_0 = 5{,}5$ тыс. руб. Доверительная вероятность $p = 99\%$.

§ 3.7.3. Доверительный интервал для индивидуальных значений переменной y при данном значении переменной x

Доверительный интервал для индивидуальных значений переменной y при заданном x_0 равен:

$$a + bx_0 \pm t_{\alpha;n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}}.$$

Пример 25. Вернемся к примеру 24. Найдем доверительный интервал для индивидуальных значений переменной y при заданном значении $x_0 = 5{,}5$ тыс. шт. Доверительная вероятность $p = 95\%$.

$$a + bx_0 \pm t_{\alpha;n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \overline{x})^2}{\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2 / n}} =$$

$$= 2{,}12 - 0{,}11 \times 5{,}5 \pm 3{,}182 \times 0{,}095 \sqrt{1 + \frac{1}{5} + \frac{(5{,}5-4)^2}{90 - 20^2/5}} \approx$$

$$\approx 1{,}515 \pm 0{,}361.$$

То есть доверительный интервал для индивидуальных значений переменной y при заданном значении $x_0 = 5{,}5$ тыс. шт. равен (1,154; 1,876). Понятно, что доверительный интервал для индивидуальных значений переменной y шире доверительного интервала для среднего значения переменной y (при заданном значении x_0).

Задача 25. В задаче 24 найти доверительный интервал для индивидуальных значений переменной y при заданном значении $x_0 = 5{,}5$ тыс. руб. Доверительная вероятность $p = 99\%$.

Глава 4

МНОЖЕСТВЕННАЯ ЛИНЕЙНАЯ РЕГРЕССИЯ

Обычно зависимую переменную называют *результативным признаком*, а независимую переменную — *фактором*. Очень часто наблюдается случай, когда результативный признак зависит не от одного, а от многих факторов. Тогда вместо парной линейной регрессии используют множественную линейную регрессию: $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_m x_m + \varepsilon$, где ε — это ошибка. Пусть n — число наблюдений, m — число объясняющих переменных. Наша задача — оценить параметры модели $\beta_0, \beta_1, \ldots, \beta_m$. Для обеспечения статистической надежности требуется выполнение условия $n \geq 3(m + 1)$.

§ 4.1. ОСНОВНЫЕ ПРЕДПОСЫЛКИ МОДЕЛИ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ

Основные предпосылки:

1) математическое ожидание случайного отклонения ε_i равно нулю для всех наблюдений;

2) дисперсии отклонений постоянны и равны для всех наблюдений;

3) случайные отклонения независимы друг от друга;

4) случайное отклонение независимо от объясняющих переменных;

5) модель линейна относительно параметров;

6) между факторами отсутствует строгая линейная связь;

7) случайные отклонения ε_i распределены нормально с параметрами 0 и σ^2: $\varepsilon_i \sim N(0, \sigma^2)$.

§ 4.2. РАСЧЕТ КОЭФФИЦИЕНТОВ МНОЖЕСТВЕННОЙ ЛИНЕЙНОЙ РЕГРЕССИИ МЕТОДОМ НАИМЕНЬШИХ КВАДРАТОВ (МНК)

Рассуждаем аналогично случаю парной линейной регрессии. Подставляем вместо переменных результаты наблюде-

ний, находим остатки и минимизируем сумму квадратов остатков. Получаем $b_0, b_1, \ldots, b_m$ — оценки параметров модели $\beta_0, \beta_1, \ldots, \beta_m$ соответственно. Ограничимся случаем $m = 2$ (случай $m > 2$ будет разобран далее с применением пакета Excel). $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

$$\overline{x}_1 = \sum x_{i1}/n, \quad \overline{x}_2 = \sum x_{i2}/n, \quad \overline{y} = \sum y_i/n.$$

$$b_1 = \frac{\sum(x_{i1}-\overline{x}_1)(y_i-\overline{y})\sum(x_{i2}-\overline{x}_2)^2}{\sum(x_{i1}-\overline{x}_1)^2\sum(x_{i2}-\overline{x}_2)^2 - (\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} -$$

$$- \frac{\sum(x_{i2}-\overline{x}_2)(y_i-\overline{y})\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)}{\sum(x_{i1}-\overline{x}_1)^2\sum(x_{i2}-\overline{x}_2)^2 - (\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2}.$$

$$b_2 = \frac{\sum(x_{i2}-\overline{x}_2)(y_i-\overline{y})\sum(x_{i1}-\overline{x}_1)^2}{\sum(x_{i1}-\overline{x}_1)^2\sum(x_{i2}-\overline{x}_2)^2 - (\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} -$$

$$- \frac{\sum(x_{i1}-\overline{x}_1)(y_i-\overline{y})\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)}{\sum(x_{i1}-\overline{x}_1)^2\sum(x_{i2}-\overline{x}_2)^2 - (\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2}.$$

$$b_0 = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2.$$

Во всех формулах суммирование от 1 до n.

Пример 26. Предполагается, что объем предложения товара y линейно зависит от цены товара x_1 и зарплаты сотрудников x_2: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$. Статистические данные собраны за 10 месяцев. Оценим по МНК коэффициенты уравнения регрессии. Заполняем таблицу.

y_i	x_{i1}	x_{i2}	$x_{i1}-\overline{x}_1$	$(x_{i1}-\overline{x}_1)^2$	$x_{i2}-\overline{x}_2$	$(x_{i2}-\overline{x}_2)^2$	$y_i-\overline{y}$
20	10	12	–22	484	4,5	20,25	–44
35	15	10	–17	289	2,5	6,25	–29
30	20	9	–12	144	1,5	2,25	–34
45	25	9	–7	49	1,5	2,25	–19
60	40	8	8	64	0,5	0,25	–4
70	37	8	5	25	0,5	0,25	6
75	43	6	11	121	–1,5	2,25	11
90	35	4	3	9	–3,5	12,25	26
105	40	4	8	64	–3,5	12,25	41
110	55	5	23	529	–2,5	6,25	46
640	320	75		1778		64,5	Сумма

$(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)$	$(x_{i1}-\overline{x}_1)(y_i-\overline{y})$	$(x_{i2}-\overline{x}_2)(y_i-\overline{y})$
−99	968	−198
−42,5	493	−72,5
−18	408	−51
−10,5	133	−28,5
4	−32	−2
2,5	30	3
−16,5	121	−16,5
−10,5	78	−91
−28	328	−143,5
−57,5	1058	−115
−276	3585	−715

$\overline{x}_1 = \sum x_{i1}/n = 320/10 = 32,$

$\overline{x}_2 = \sum x_{i2}/n = 75/10 = 7{,}5,$

$\overline{y} = \sum y_i/n = 640/10 = 64.$

$$b_1 = \frac{\sum(x_{i1}-\overline{x}_1)(y_i-\overline{y})\sum(x_{i2}-\overline{x}_2)^2}{\sum(x_{i1}-\overline{x}_1)^2\sum(x_{i2}-\overline{x}_2)^2-(\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} -$$

$$- \frac{\sum(x_{i2}-\overline{x}_2)(y_i-\overline{y})\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)}{\sum(x_{i1}-\overline{x}_1)^2\sum(x_{i2}-\overline{x}_2)^2-(\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} =$$

$$= \frac{3585\times 64{,}5-(-715)\times(-276)}{1778\times 64{,}5-(-276)^2} \approx 0{,}88.$$

$$b_2 = \frac{\sum(x_{i2}-\overline{x}_2)(y_i-\overline{y})\sum(x_{i1}-\overline{x}_1)^2}{\sum(x_{i1}-\overline{x}_1)^2\sum(x_{i2}-\overline{x}_2)^2-(\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} -$$

$$- \frac{\sum(x_{i1}-\overline{x}_1)(y_i-\overline{y})\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)}{\sum(x_{i1}-\overline{x}_1)^2\sum(x_{i2}-\overline{x}_2)^2-(\sum(x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2}.$$

$$= \frac{(-715)\times 1778-3585\times(-276)}{1778\times 64{,}5-(-276)^2} \approx -7{,}32.$$

$$b_0 = \overline{y} - b_1\overline{x}_1 - b_2\overline{x}_2 = 64 - 0{,}88\times 32 - (-7{,}32)\times 7{,}5 = 90{,}74.$$

$$y = b_0 + b_1x_1 + b_2x_2 = 90{,}74 + 0{,}88x_1 - 7{,}32x_2.$$

Задача 26. Предполагается, что объем предложения товара y линейно зависит от цены товара x_1 и зарплаты сотрудников x_2: $y = \beta_0 + \beta_1x_1 + \beta_2x_2$. Статистические данные собраны за 10 месяцев. Оценить по МНК коэффициенты уравнения регрессии.

y	75	90	105	110	120	130	130	130	135	140
x_1	43	35	38	55	50	35	40	55	45	65
x_2	6	4	4	5	3	1	2	3	1	2

§ 4.3. СТАНДАРТНЫЕ ОШИБКИ КОЭФФИЦИЕНТОВ

Знание дисперсий и стандартных ошибок позволяет анализировать точность оценок, строить доверительные интервалы для теоретических коэффициентов, проверять гипотезы.

Подставив значения факторов в найденное уравнение регрессии $y = b_0 + b_1x_1 + \ldots + b_mx_m$, получим числа $\hat{y}_i$, $i = 1, \ldots, n$. Тогда разность между наблюдаемым значением y_i и расчетным значением $\hat{y}_i$ есть величина ошибки $e_i = y_i - \hat{y}_i$, $i = 1, \ldots, n$.

Стандартная ошибка регрессии $S = \sqrt{\dfrac{\sum_{i=1}^{n} e_i^2}{n-m-1}}$.

Зная S, можно найти стандартные ошибки коэффициентов. Для случая $m = 2$

$$S_{b_0}^2 = S^2\Big(\frac{\bar{x}_1^2\sum(x_{i2}-\bar{x}_2)^2 + \bar{x}_2^2\sum(x_{i1}-\bar{x}_1)^2}{\sum(x_{i1}-\bar{x}_1)^2\sum(x_{i2}-\bar{x}_2)^2 - (\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2} -$$

$$-\frac{2\bar{x}_1\bar{x}_2\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2)}{\sum(x_{i1}-\bar{x}_1)^2\sum(x_{i2}-\bar{x}_2)^2 - (\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2} + \frac{1}{n}\Big).$$

$$S_{b_1}^2 = S^2\frac{\sum(x_{i2}-\bar{x}_2)^2}{\sum(x_{i1}-\bar{x}_1)^2\sum(x_{i2}-\bar{x}_2)^2 - (\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2}.$$

$$S_{b_2}^2 = S^2\frac{\sum(x_{i1}-\bar{x}_1)^2}{\sum(x_{i1}-\bar{x}_1)^2\sum(x_{i2}-\bar{x}_2)^2 - (\sum(x_{i1}-\bar{x}_1)(x_{i2}-\bar{x}_2))^2}.$$

Во всех формулах суммирование от 1 до n.

Пример 27. Найдем стандартную ошибку регрессии и стандартные ошибки коэффициентов в примере 26. Заполняем таблицу.

y_i	x_1	x_2	$\hat{y}_i = 90{,}74 + 0{,}88x_1 - 7{,}32x_2$	$e_i = y_i - \hat{y}_i$	e_i^2
20	10	12	11,7	8,3	68,89
35	15	10	30,74	4,26	18,15
30	20	9	42,46	−12,46	155,25
45	25	9	46,86	−1,86	3,46
60	40	8	67,38	−7,38	54,46
70	37	8	64,74	5,26	27,67
75	43	6	84,66	−9,66	93,32
90	35	4	92,26	−2,26	5,11
105	40	4	96,66	8,34	69,56
110	55	5	102,54	7,46	55,65
640	320	75	Сумма		551,52

В последнем столбце результат округляем до двух цифр после запятой.

$$S^2 = \frac{\sum_{i=1}^{n} e_i^2}{n-m-1} = \frac{551,52}{10-2-1} \approx 78,79. \quad S = \sqrt{78,79} \approx 8,88.$$

$$S_{b_0}^2 = S^2 \Big(\frac{\overline{x}_1^{\,2} \sum (x_{i2}-\overline{x}_2)^2 + \overline{x}_2^{\,2} \sum (x_{i1}-\overline{x}_1)^2}{\sum (x_{i1}-\overline{x}_1)^2 \sum (x_{i2}-\overline{x}_2)^2 - (\sum (x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} -$$

$$- \frac{2\overline{x}_1\overline{x}_2 \sum (x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2)}{\sum (x_{i1}-\overline{x}_1)^2 \sum (x_{i2}-\overline{x}_2)^2 - (\sum (x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} + \frac{1}{n} \Big) =$$

$$= 78,79 \Big(\frac{32^2 \times 64,5 + 7,5^2 \times 1778 - 2 \times 32 \times 7,5 \times (-276)}{1778 \times 64,5 - (-276)^2} +$$

$$+ \frac{1}{10} \Big) \approx 618,76. \text{ Отсюда } S_{b_0} = 24,87.$$

$$S_{b_1}^2 = S^2 \frac{\sum (x_{i2}-\overline{x}_2)^2}{\sum (x_{i1}-\overline{x}_1)^2 \sum (x_{i2}-\overline{x}_2)^2 - (\sum (x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} =$$

$$= 78,79 \, \frac{64,5}{38505} \approx 0,1320. \text{ Отсюда } S_{b_1} = 0,36.$$

$$S_{b_2}^2 = S^2 \frac{\sum (x_{i1}-\overline{x}_1)^2}{\sum (x_{i1}-\overline{x}_1)^2 \sum (x_{i2}-\overline{x}_2)^2 - (\sum (x_{i1}-\overline{x}_1)(x_{i2}-\overline{x}_2))^2} =$$

$$= 78,79 \, \frac{1778}{38505} \approx 3,6381. \text{ Отсюда } S_{b_2} = 1,91.$$

Задача 27. Найти стандартную ошибку регрессии и стандартные ошибки коэффициентов в задаче 26.

§ 4.4. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ТЕОРЕТИЧЕСКОГО УРАВНЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Задается доверительная вероятность p. По найденным точечным оценкам b_i коэффициентов β_i, $i = 0, 1, \ldots, m$, можно построить p-процентные доверительные интервалы. $\alpha = (1-p)/2$. Из таблиц t-распределения находим $t_{\alpha;n-m-1}$. Тогда p-процентный доверительный интервал коэффициента β_i задается формулой $b_i \pm t_{\alpha;n-m-1} \times S_{b_i}$, $i = 0, 1, \ldots, m$.

Пример 28. Найдем доверительные интервалы коэффициентов теоретического уравнения линейной регрессии в примерах 26 и 27. Доверительная вероятность 95%.

$p = 0,95$. $\alpha = (1 - p)/2 = (1 - 0,95)/2 = 0,025$. $n = 10$, $m = 2$. Из таблиц t-распределения находим $t_{\alpha;n-m-1} = t_{0,025;10-2-1} = 2,365$.

$b_0 \pm t_{\alpha;n-m-1} \times S_{b_0} = 90,74 \pm 2,365 \times 24,87 \approx 90,74 \pm 58,82$, то есть $31,92 < \beta_0 < 149,56$.

$b_1 \pm t_{\alpha;n-m-1} \times S_{b_1} = 0,88 \pm 2,365 \times 0,36 \approx 0,88 \pm 0,85$, то есть $0,03 < \beta_1 < 1,73$.

$b_2 \pm t_{\alpha;n-m-1} \times S_{b_2} = -7,32 \pm 2,365 \times 1,91 \approx -7,32 \pm 4,52$, то есть $-11,84 < \beta_2 < -2,80$.

Задача 28. Найти доверительные интервалы коэффициентов теоретического уравнения линейной регрессии в задачах 26 и 27. Доверительная вероятность 99%.

§ 4.5. ПРОВЕРКА СТАТИСТИЧЕСКОЙ ЗНАЧИМОСТИ КОЭФФИЦИЕНТОВ УРАВНЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

Мы включили в модель m объясняющих переменных $x_1, \ldots, x_m$. Возможно, что не все из них влияют на результативный признак y. Поэтому проводится проверка статистической значимости коэффициентов уравнения линейной регрессии.

H_0: $\beta_i = 0$, то есть объясняющая переменная x_i не влияет на результативный признак y.

H_1: $\beta_i \neq 0$, то есть объясняющая переменная x_i влияет на результативный признак y.

Доверительная вероятность p. $\alpha = (1 - p)/2$. Граничные точки $\pm t_{\alpha;n-m-1}$. Статистика $t_{b_i} = b_i/S_{b_i}$.

Пример 29. Определим статистическую значимость коэффициентов теоретического уравнения линейной регрессии в примерах 26–28. Доверительная вероятность 95%.

H_0: $\beta_i = 0$, то есть объясняющая переменная x_i не влияет на результативный признак y.

H_1: $\beta_i \neq 0$, то есть объясняющая переменная x_i влияет на результативный признак y.

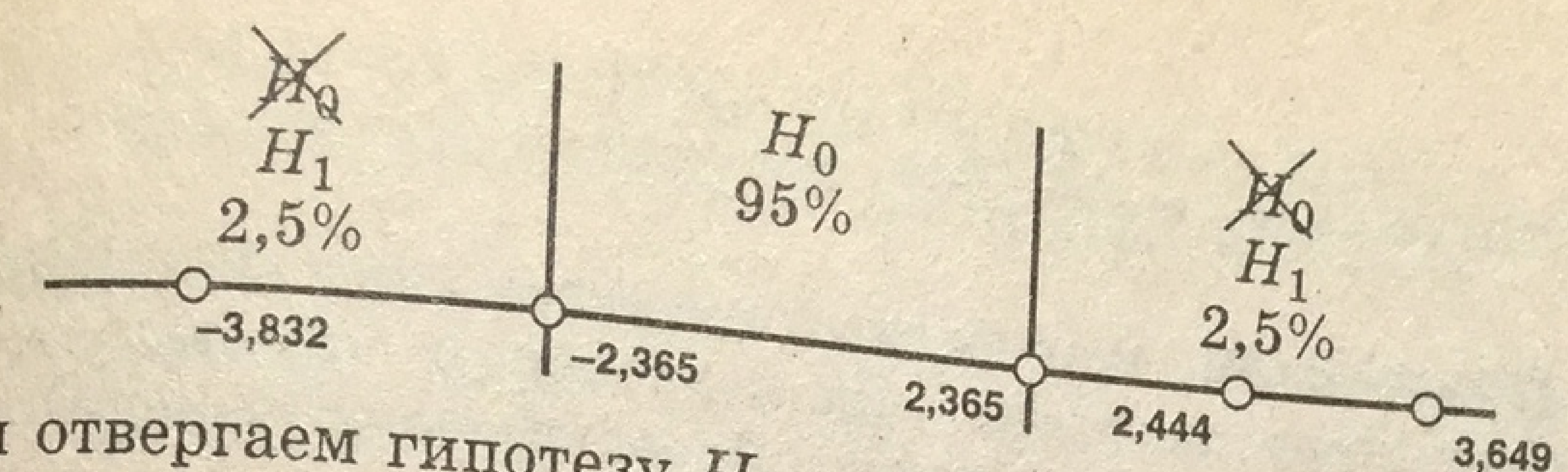
Проведем двустороннюю проверку. Граничные точки $\pm t_{\alpha;n-m-1} = \pm 2,365$. Статистики:

$t_{b_0} = b_0/S_{b_0} = 90,74/24,87 \approx 3,649$,

$t_{b_1} = b_1/S_{b_1} = 0,88/0,36 \approx 2,444$,

$t_{b_2} = b_2/S_{b_2} = -7,32/1,91 \approx -3,832$.

Отметим значения на числовой оси.



Мы отвергаем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Все коэффициенты статистически значимы, то есть x_1 и x_2 влияют на результативный признак y.

Задача 29. Определить статистическую значимость коэффициентов теоретического уравнения линейной регрессии в задачах 26–28. Доверительная вероятность 99%.

§ 4.6. ПРОВЕРКА ОБЩЕГО КАЧЕСТВА УРАВНЕНИЯ ЛИНЕЙНОЙ РЕГРЕССИИ

После исследования статистической значимости коэффициентов уравнения линейной регрессии проверяют общее качество уравнения линейной регрессии. С этой целью вычисляют *коэффициент детерминации*:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2}.$$

Коэффициент детерминации — это доля общего разброса значений результативного признака y, объясненная уравнением линейной регрессии. $0 \le R^2 \le 1$. Чем ближе R^2 к 1, тем лучше полученное уравнение линейной регрессии объясняет поведение результативного признака y. Величина R^2 не убывает с введением в модель дополнительного фактора. Исправленный коэффициент детерминации

$$\overline{R}^2 = 1 - \frac{(1 - R^2)(n-1)}{n-m-1}$$

с введением в модель дополнительного фактора растет медленнее, чем R^2.

С целью анализа совокупной значимости коэффициентов уравнения линейной регрессии проверяют гипотезу о статистической значимости коэффициента детерминации.

H_0: $R^2 = 0$.
H_1: $R^2 > 0$.

Доверительная вероятность p. Правосторонняя проверка. $\alpha = 1 - p$. Из таблиц F-распределения находим граничную точку $F_{\alpha;m;n-m-1}$.

Статистика $F = \frac{R^2}{1-R^2} \times \frac{n-m-1}{m}$.

Пример 30. Вернемся к примерам 26 и 27. Найдем коэффициент детерминации и проверим гипотезу о его статистической значимости. Доверительная вероятность $p = 95\%$.

$y_i - \overline{y}$	−44	−29	−34	−19	−4	6	11	26	41	46	Сумма
$(y_i - \overline{y})^2$	1936	841	1156	361	16	36	121	676	1681	2116	8940

Коэффициент детерминации:

$$R^2 = 1 - \frac{\sum e_i^2}{\sum (y_i - \overline{y})^2} = 1 - \frac{551{,}52}{8940} \approx 0{,}938.$$

То есть наша модель объясняет 93,8% общего разброса значений результативного признака y.

Исправленный коэффициент детерминации:

$$\overline{R}^2 = 1 - \frac{(1-R^2)(n-1)}{n-m-1} = 1 - \frac{(1-0{,}938)(10-1)}{10-2-1} \approx 0{,}920.$$

H_0: $R^2 = 0$.
H_1: $R^2 > 0$.

Доверительная вероятность $p = 0{,}95$. Правосторонняя проверка. $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$.

Из таблиц F-распределения находим граничную точку $F_{\alpha;m;n-m-1} = F_{0,05;2;10-2-1} = 4{,}74$. Статистика:

$$F = \frac{R^2}{1-R^2} \times \frac{n-m-1}{m} = \frac{0{,}938}{1-0{,}938} \times \frac{10-2-1}{2} \approx 52{,}95.$$

Отметим значения на числовой оси.

H0
95%
~~H0~~
H1
5%
4,74
52,9

Мы отвергаем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Предположение о незначительности связи отвергается.

Задача 30. В задачах 26 и 27 найти коэффициент детерминации и проверить гипотезу о его статистической значимости. Доверительная вероятность $p = 99\%$.

§ 4.7. ПРОВЕРКА РАВЕНСТВА ДВУХ КОЭФФИЦИЕНТОВ ДЕТЕРМИНАЦИИ

По n наблюдениям построено уравнение линейной регрессии, содержащее m факторов. Для этой модели коэффициент детерминации R_1^2. После этого из модели исключили k объясняющих переменных. Для нового уравнения линейной регрессии коэффициент детерминации R_2^2. Существенно ли ухудшилось качество описания поведения результативного признака y? Ответ на этот вопрос дает следующая проверка гипотезы.

H_0: $R_1^2 = R_2^2$, то есть качество описания поведения результативного признака y существенно не ухудшилось.

H_1: $R_1^2 > R_2^2$, то есть качество описания поведения результативного признака y ухудшилось существенно.

Доверительная вероятность p. Правосторонняя проверка. $\alpha = 1 - p$. Из таблиц F-распределения находим граничную точку $F_{\alpha;k;n-m-1}$.

Статистика $F = \dfrac{R_1^2 - R_2^2}{1 - R_1^2} \times \dfrac{n-m-1}{k}$.

Пример 31. По $n = 15$ наблюдениям построено уравнение линейной регрессии, содержащее $m = 4$ фактора. Для этой модели коэффициент детерминации $R_1^2 = 0{,}95$. После этого из модели исключили $k = 1$ объясняющую переменную. Для нового уравнения линейной регрессии коэффициент детерминации $R_2^2 = 0{,}9$. Существенно ли ухудшилось качество описания поведения результативного признака y? Доверительная вероятность $p = 95\%$.

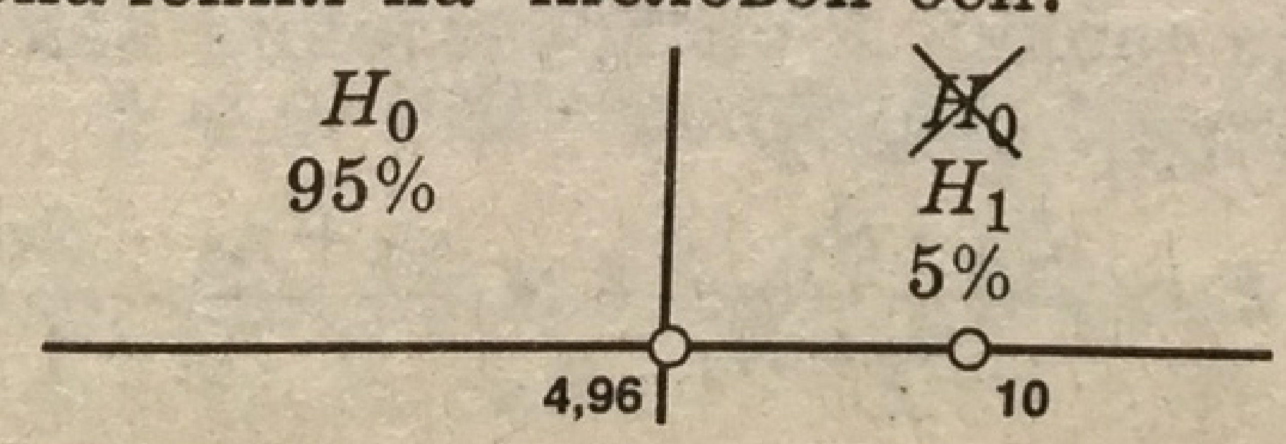
H_0: $R_1^2 = R_2^2$, то есть качество описания поведения результативного признака y существенно не ухудшилось.

H_1: $R_1^2 > R_2^2$, то есть качество описания поведения результативного признака y ухудшилось существенно.

Правосторонняя проверка. $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$. Из таблиц F-распределения находим граничную точку $F_{\alpha;k;n-m-1} = F_{0{,}05;1;15-4-1} = 4{,}96$. Статистика:

$$F = \frac{R_1^2 - R_2^2}{1 - R_1^2} \times \frac{n-m-1}{k} = \frac{0{,}95 - 0{,}9}{1 - 0{,}95} \times \frac{15-4-1}{1} = 10.$$

Отметим значения на числовой оси.



Мы отвергаем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Качество описания поведения результативного признака y ухудшилось существенно.

Задача 31. По $n = 12$ наблюдениям построено уравнение линейной регрессии, содержащее $m = 3$ фактора. Для этой модели коэффициент детерминации $R_1^2 = 0{,}90$. После этого из модели исключили $k = 2$ объясняющих переменных. Для нового уравнения линейной регрессии коэффициент детерминации $R_2^2 = 0{,}84$. Существенно ли ухудшилось качество описания поведения результативного признака y? Доверительная вероятность $p = 99\%$.

Замечание. Аналогичные рассуждения могут быть использованы для выяснения обоснованности включения в модель новых k объясняющих переменных. В этом случае H_1: $R_2^2 > R_1^2$.

Статистика $F = \dfrac{R_2^2 - R_1^2}{1 - R_2^2} \times \dfrac{n - m - 1}{k}$.

§ 4.8. ПРОВЕРКА ГИПОТЕЗЫ О СОВПАДЕНИИ УРАВНЕНИЙ РЕГРЕССИИ ДЛЯ ДВУХ ВЫБОРОК. ТЕСТ ЧОУ

Производится выборка объема n_1. По ней строится уравнение регрессии $y = b_{01} + b_{11}x_1 + \dots + b_{m1}x_m + e_1$. Производится выборка объема n_2. По ней строится уравнение регрессии $y = b_{02} + b_{12}x_1 + \dots + b_{m2}x_m + e_2$. Будет ли уравнение регрессии одним и тем же для обеих выборок?

Для каждой выборки находим сумму квадратов остатков $S_1 = \sum_{i=1}^{n} e_{i1}^2$ и $S_2 = \sum_{i=1}^{n} e_{i2}^2$ соответственно. По объединенной выборке объема $n_1 + n_2$ строим уравнение регрессии, для которого также находим сумму квадратов остатков $S_0 = \sum_{i=1}^{n} e_i^2$.

H_0: $b_{i1} = b_{i2}$, $i = 0, 1, \dots, m$, то есть коэффициенты уравнений линейной регрессии одинаковы.

H_1: коэффициенты уравнений линейной регрессии различны.

Доверительная вероятность p. Правосторонняя проверка. $\alpha = 1 - p$. Из таблиц F-распределения находим граничную точку $F_{\alpha;m+1;n_1+n_2-2m-2}$.

Статистика $F = \dfrac{S_0 - S_1 - S_2}{S_1 + S_2} \times \dfrac{n_1 + n_1 - 2m - 2}{m + 1}$.

Пример 32. По $n = 25$ наблюдениям построено уравнение линейной регрессии, содержащее $m = 2$ фактора. Есть основания предполагать, что модель будет более реалистичной, если весь интервал наблюдений разбить на два подынтервала и оценивать уравнение линейной регрессии для каждого из них отдельно. Это связано с изменением институциональных условий между 10-м и 11-м наблюдениями. Суммы квадратов остатков для общей выборки $S_0 = 140$, для 1-го подынтервала $S_1 = 100$, для 2-го подынтервала $S_2 = 30$. Есть ли основания считать, что это разбиение целесообразно? Доверительная вероятность $p = 95\%$.

H_0: $b_{i1} = b_{i2}$, $i = 0, 1, \ldots, m$, то есть коэффициенты уравнений линейной регрессии одинаковы.

H_1: коэффициенты уравнений линейной регрессии различны.

Доверительная вероятность $p = 0{,}95$. Правосторонняя проверка. $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$. $n_1 = 10$, $n_2 = 25 - 10 = 15$. Из таблиц F-распределения находим граничную точку $F_{\alpha;m+1;n_1+n_2-2m-2} = F_{0,05;2+1;10+15-2\times2-2} = 3{,}13$.

$$\text{Статистика } F = \frac{S_0 - S_1 - S_2}{S_1 + S_2} \times \frac{n_1 + n_1 - 2m - 2}{m + 1} =$$

$$= \frac{140 - 100 - 30}{100 + 30} \times \frac{10 + 15 - 2\times 2 - 2}{2 + 1} \approx 0{,}49 < 3{,}13.$$

Мы принимаем гипотезу H_0 на уровне значимости 5%. Для всего рассматриваемого периода нужно строить единое уравнение регрессии.

Задача 32. По $n = 24$ наблюдениям построено уравнение линейной регрессии, содержащее $m = 2$ фактора. Есть основания предполагать, что модель будет более реалистичной, если весь интервал наблюдений разбить на два подынтервала и оценивать уравнение линейной регрессии для каждого из них отдельно. Это связано с изменением институциональных условий между 12-м и 13-м наблюдениями. Суммы квадратов остатков для общей выборки $S_0 = 120$, для 1-го подынтервала $S_1 = 80$, для 2-го подынтервала $S_2 = 25$. Есть ли основания считать, что это разбиение целесообразно? Доверительная вероятность $p = 99\%$.

§ 4.9. РЕГРЕССИЯ И Excel

Excel позволяет при построении уравнения линейной регрессии большую часть работы сделать очень быстро. Важ-

но понять, как интерпретировать полученные результаты. Воспользуемся надстройкой *Пакет анализа.*

Сервис → *Анализ данных* → *Регрессия* → *ОК.* Появляется диалоговое окно, которое нужно заполнить. В графе *Входной интервал Y:* указывается ссылка на ячейки, содержащие значения результативного признака y. В графе *Входной интервал X:* указывается ссылка на ячейки, содержащие значения факторов $x_1, \dots, x_m$ ($m \leq 16$). *Уровень надежности* (доверительная вероятность) по умолчанию предполагается равным 95%. Если исследователя это значение не устраивает, то рядом со словами *Уровень надежности* нужно поставить «галочку» и указать требуемое значение. Поставив «галочку» рядом со словом *константа-ноль*, исследователь получит $b_0 = 0$ по умолчанию. Если нужны значения остатков e_i и их график, то нужно поста-

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	R
R-квадрат	R^2
Нормированный R-квадрат	$\bar{R}^2$
Стандартная ошибка	S
Наблюдения	n

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	m	$\Sigma(\hat{y}_i - \bar{y})^2$	SS/df	Статистика F= = MS(регр)/ MS(ост)	Fрасп (F; df(регр); df(ост))
Остаток	n−m−1	$\Sigma(y_i - \hat{y}_i)^2$	SS/df		
Итого	n − 1	Сумма			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-значение*	*Нижние 95%*	*Верхние 95%*	*Нижние*	*Верхние*
y-пересечение	b_0	S_{b_0}	t_{b_0}					
x_1	b_1	S_{b_1}	t_{b_1}					
x_2	b_2	S_{b_2}	t_{b_2}					

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанный y*	*Остатки*
Номер	$\hat{y}_i$	e_i

вить «галочки» рядом со словами *Остатки* и *График остатков. OK*. Появляется итоговое окно.

Если число в графе *Значимость F* превышает 1 – *Уровень надежности*, то принимается гипотеза $R^2 = 0$. Иначе принимается гипотеза $R^2 \neq 0$.

P-значение — это значения уровней значимости, соответствующие вычисленным t-статистикам. P-значение = = СТЬЮДРАСП (t-статистика; $n - m - 1$) (статистическая функция мастера функций f_x). Если P-значение превышает 1 – *Уровень надежности*, то соответствующая переменная статистически незначима и ее можно исключить из модели.

Нижние 95% и *Верхние 95%* — это нижние и верхние границы 95-процентных доверительных интервалов для коэффициентов теоретического уравнения линейной регрессии. Если исследователь согласился с принятым по умолчанию значением доверительной вероятности 95%, то последние два столбца будут дублировать два предыдущих столбца. Если исследователь вводил свое значение доверительной вероятности p, то последние два столбца содержат значения соответственно нижней и верхней границы p-процентных доверительных интервалов.

Если же надстройки *Пакет анализа* нет, то можно воспользоваться статистической функцией ЛИНЕЙН мастера функций f_x. Перед вызовом это функции нужно выделить диапазон ячеек следующего размера (для парной регрессии это блок размера 5×2).

b_m	b_{m-1}	...	b_1	b_0
S_{b_m}	$S_{b_{m-1}}$	...	S_{b_1}	S_{b_0}
R^2	S			
Статистика F	n–m–1			
$\Sigma(\hat{y}_i - \overline{y})^2$	$\Sigma(y_i - \hat{y}_i)^2$			

Тогда после выполнения процедуры в ячейках будут находиться указанные величины. *fx → статистические → ЛИНЕЙН → OK*. Появляется диалоговое окно, которое нужно заполнить. Если исследователю требуется $b_0 = 0$, то в графе *константа* нужно ввести значение 0. В графе *статистика* указывается значение 1. После этого нажимается не *OK*, а комбинация клавиш *Ctrl + Shift + Enter*.

Глава 5

ГЕТЕРОСКЕДАСТИЧНОСТЬ

При использовании МНК и в парной, и во множественной линейных регрессиях были наложены некоторые ограничения. Ближайшие три главы будут посвящены изучению выполнимости предпосылок МНК.

Одной из предпосылок МНК является условие постоянства дисперсий случайных отклонений (*гомоскедастичность*). Не должно быть априорной причины, вызывающей большую ошибку (отклонение) при одних наблюдениях и меньшую — при других. Невыполнимость данной предпосылки называется *гетероскедастичностью*.

На практике гетероскедастичность не так уж и редка. Проблема гетероскедастичности характерна для перекрестных данных и довольно редко встречается при рассмотрении временных рядов. Оценки, полученные по МНК, при наличии гетероскедастичности не будут эффективными (то есть они не будут иметь наименьшую дисперсию по сравнению с другими оценками данного параметра). Стандартные ошибки коэффициентов S_{b_i} будут занижены. Поэтому статистики $t_{b_i} = b_i/S_{b_i}$ будут завышены, что может привести к признанию статистически значимыми коэффициентов, которые таковыми не являются. Доверительные интервалы $b_i \pm t_{\alpha;n-m-1}\, S_{b_i}$ теоретических коэффициентов уравнения линейной регрессии получаются шире, чем на самом деле.

Как выяснить наличие гетероскедастичности и смягчить ее последствия? Не существует однозначного метода определения гетероскедастичности.

§ 5.1. ТЕСТ РАНГОВОЙ КОРРЕЛЯЦИИ СПИРМЕНА

Предполагается, что дисперсии отклонений будут либо увеличиваться, либо уменьшаться с ростом значений x. Пусть n — число наблюдений. Значения переменной x и $|e_i|$ ран-

жируются (упорядочиваются по величине). Обозначим через d разность между рангами значений переменной x и $|e_i|$.

Коэффициент ранговой корреляции $r_{x,e} = 1 - 6\dfrac{\sum d^2}{n(n^2-1)}$.

Зададим доверительную вероятность p. $\alpha = (1-p)/2$. По t-таблицам находим граничную точку $t_{\alpha;n-2}$.

Статистика $t = \dfrac{r_{x,e}\sqrt{n-2}}{\sqrt{1-r_{x,e}^2}}$.

Если $t < t_{\alpha;n-2}$, то на уровне значимости α принимается гипотеза об отсутствии гетероскедастичности. Иначе гипотеза об отсутствии гетероскедастичности отклоняется. В модели, содержащей несколько факторов, проверка гипотезы об отсутствии гетероскедастичности проводится с помощью статистики t для каждого из них отдельно.

Пример 33. В примерах 18 и 19 проверим гипотезу об отсутствии гетероскедастичности с помощью теста ранговой корреляции Спирмена. Доверительная вероятность $p = 95\%$.

x	e_i	$\lvert e_i \rvert$	d_1	d_2	$d=d_1-d_2$	d^2
2	0	0	5	5	0	0
3	–0,09	0,09	4	2	2	4
4	0,12	0,12	3	1	2	4
5	0,03	0,03	2	4	–2	4
6	–0,06	0,06	1	3	–2	4
Сумма						16

Заполним таблицу. Модули элементов второго столбца запишем в 3-й столбец. В 4-м и 5-м столбцах ранжированы по убыванию элементы 1-го и 3-го столбцов соответственно. $n = 5$ наблюдений.

$$r_{x,e} = 1 - 6\frac{\sum d^2}{n(n^2-1)} = 1 - 6\frac{16}{5(5^2-1)} = 0{,}2.$$

$\alpha = (1-p)/2 = (1-0{,}95)/2 = 0{,}025$. По t-таблицам находим граничную точку $t_{\alpha;n-2} = t_{0,025;5-2} = 3{,}182$.

Статистика $t = \dfrac{r_{x,e}\sqrt{n-2}}{\sqrt{1-r_{x,e}^2}} = \dfrac{0{,}2\sqrt{5-2}}{\sqrt{1-0{,}2^2}} \approx 0{,}354 < 3{,}182$.

Мы принимаем гипотезу об отсутствии гетероскедастичности на уровне значимости 5%.

Задача 33. В задачах 18 и 19 проверить гипотезу об отсутствии гетероскедастичности с помощью теста ранговой корреляции Спирмена. Доверительная вероятность $p = 99\%$.

Замечание. При ранжировании значений переменной можно воспользоваться надстройкой *Пакет анализа* в Excel. *Сервис → Анализ данных → Ранг и персентиль → OK*. Появляется диалоговое окно, которое нужно заполнить. Выбирается способ группирования данных (*по строкам* или *по столбцам*). *OK*. Данные упорядочиваются по убыванию.

При отсутствии надстройки *Пакет анализа* можно воспользоваться статистической функцией РАНГ (число; ссылка; порядок) мастера функций f_x пакета Excel. Здесь *число* — это значение переменной, для которого определяется ранг в наборе данных. *Ссылка* — это ссылка на ячейки, которые содержат значения переменной. Если *порядок* = 0 или опущен, то определяется ранг значения переменной для данных, упорядоченных по убыванию. Если *порядок* = 1, то определяется ранг значения переменной для данных, упорядоченных по возрастанию.

§ 5.2. ТЕСТ ГОЛДФЕЛДА-КВАНДТА

Рассматривается связь величин вида $y = a + bx$. Предполагается, что стандартное отклонение $\sigma_i = \sigma(\varepsilon_i)$ пропорционально значению переменной x в этом наблюдении: $\sigma_i^2 = \sigma^2 x_i^2$, $i = 1, \ldots, n$, n — число наблюдений. Также предполагается, что ε_i имеет нормальное распределение и отсутствует автокорреляция (будет рассмотрена в дальнейшем). Все n наблюдений упорядочиваются по величине x. Эта упорядоченная выборка делится на три примерно равные части объемов k, $n - 2k$ и k соответственно. При $n = 30$ $k = 11$, при $n = 60$ $k = 22$.

Для каждой из выборок объема k оценивается свое уравнение регрессии и находятся суммы квадратов отклонений $S_1 = \sum_{i=1}^{k} e_i^2$ и $S_3 = \sum_{i=n-k+1}^{n} e_i^2$ соответственно.

Зададим доверительную вероятность p. $\alpha = 1 - p$. По F-таблицам находим граничную точку $F_{\alpha;k-m-1;k-m-1}$, где m — число факторов модели. Статистика $F = S_3/S_1$.

Если $F < F_{\alpha;k-m-1;k-m-1}$, то на уровне значимости α принимается гипотеза об отсутствии гетероскедастичности.

Иначе гипотеза об отсутствии гетероскедастичности отклоняется. Для множественной регрессии тест обычно проводится для того фактора, который в максимальной степени связан с σ_i. При этом выбирают $k > m + 1$. Если нет уверенности относительно выбора фактора x_j, то данный тест можно осуществить для каждого фактора.

Пример 34. Рассматривается регрессионная линейная модель с $m = 2$ факторами. $n = 30$ наблюдений. Для первых и последних $k = 11$ наблюдений суммы квадратов отклонений $S_1 = 20$ и $S_3 = 45$ соответственно. С помощью теста Голдфелда-Квандта проверим гипотезу об отсутствии гетероскедастичности.

Зададим доверительную вероятность $p = 95\%$.

$\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$. По F-таблицам находим граничную точку $F_{\alpha;k-m-1;k-m-1} = F_{0,05;11-2-1;11-2-1} = 3{,}44$.

Статистика $F = S_3/S_1 = 45/20 = 2{,}25 < 3{,}44$. На уровне значимости 5% принимается гипотеза об отсутствии гетероскедастичности.

Задача 34. Рассматривается регрессионная линейная модель с $m = 2$ факторами. $n = 30$ наблюдений. Для первых и последних $k = 11$ наблюдений суммы квадратов отклонений $S_1 = 18$ и $S_3 = 52$ соответственно. С помощью теста Голдфелда-Квандта проверить гипотезу об отсутствии гетероскедастичности. Доверительная вероятность $p = 99\%$.

§ 5.3. СМЯГЧЕНИЕ ПРОБЛЕМЫ ГЕТЕРОСКЕДАСТИЧНОСТИ. МЕТОД ВЗВЕШЕННЫХ НАИМЕНЬШИХ КВАДРАТОВ (ВНК)

После установления гетероскедастичности модель преобразуют с целью устранения этого недостатка. Вид преобразования зависит от того, известны или нет дисперсии σ_i^2 отклонений ε_i, $i = 1, \ldots, n$. Используют метод ВНК. Ограничимся случаем, когда $y = \beta_0 + \beta_1 x + \varepsilon$, дисперсии σ_i^2 неизвестны и пропорциональны x_i^2.

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \Rightarrow y_i/x_i = \beta_0/x_i + \beta_1 + \varepsilon_i/x_i \Rightarrow$$
$$\Rightarrow y_i/x_i = \beta_0/x_i + \beta_1 + v_i, \text{ где } v_i = \varepsilon_i/x_i.$$

Для этого уравнения уже выполняется условие гомоскедастичности. По МНК находим оценки коэффициентов β_0, β_1 и возвращаемся к исходному уравнению. В случае, когда число факторов $m > 1$, исходное уравнение делится на переменную, которая в максимальной степени связана с σ_i.

Пример 35. Для предприятий отрасли анализируется зарплата y в зависимости от x (количество сотрудников). $n = 30$ случайно отобранных предприятий.

x	y					
100	75,5	75,5	77,5	78,5	80	81
200	80,5	82	84,5	85	85,5	86,5
300	85,5	88,5	90	91	95	96
400	93	93,5	97,5	99	102,5	105
500	102	105,5	107	110,5	115	118,5

Уравнение линейной регрессии $y = \beta_0 + \beta_1 x + \varepsilon$. Так как с ростом x разброс значений y увеличивается, то можно ожидать наличие гетероскедастичности. Предположим, что гетероскедастичность имеет место и σ_i^2 пропорциональны x_i^2. Введем новые обозначения $z = y/x$ и $t = 1/x$ и перейдем к уравнению $z = \beta_0 t + \beta_1$. По МНК находим оценки коэффициентов β_0, β_1 и возвращаемся к исходному уравнению $y = \beta_0 + \beta_1 x$.

Задача 35. В примере 35 с помощью теста Голдфелда-Квандта проверить гипотезу об отсутствии гетероскедастичности ($k = 12$). Доверительная вероятность $p = 95\%$. Устранить гетероскедастичность и найти оценки коэффициентов β_0, β_1. Рекомендуется воспользоваться пакетом Excel.

Глава 6

АВТОКОРРЕЛЯЦИЯ

Одна из предпосылок МНК — это независимость значений случайных отклонений от значений отклонений во всех других наблюдениях. *Автокорреляция* (последовательная корреляция) — это корреляция между наблюдаемыми показателями, упорядоченными во времени (временные ряды) или в пространстве (перекрестные ряды). Автокорреляция остатков (отклонений) обычно встречается при использовании данных временных рядов. Последствия автокорреляции в определенной степени сходны с последствиями гетероскедастичности.

Нам неизвестны истинные значения отклонений ε_i, $i = 1, ..., n$. Поэтому выводы об их независимости делаются на основе оценок e_i, $i = 1, ..., n$. При этом обычно проверяется их некоррелированность, являющаяся необходимым, но недостаточным условием независимости. Проверяется некоррелированность только соседних величин e_i. При наличии автокорреляции остатков полученное уравнение регрессии считается неудовлетворительным.

Рассмотрим возможные методы определения автокорреляции и способы ее устранения.

§ 6.1. МЕТОД РЯДОВ

Определяются знаки отклонений e_i. Ряд — это непрерывная последовательность одинаковых знаков. Длина ряда — это количество знаков в ряду. n — число наблюдений, n_1 — общее количество знаков «+», n_2 — общее количество знаков «–», k — количество рядов. Доверительная вероятность $p = 0{,}95$, уровень значимости $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05$.

Для небольшого числа наблюдений ($n_1 < 20$, $n_2 < 20$) из специальных таблиц Сведа-Эйзенхарта по n_1, n_2 и $\alpha = 0{,}05$ находят числа k_1 и k_2. Если $k_1 < k < k_2$, то автокорреляция отсутствует. Если $k \leq k_1$, то говорят о *положительной автокорреляции остатков*. Если $k \geq k_2$, то говорят об *от-*

рицательной автокорреляции остатков (за положительным отклонением следует отрицательное и наоборот).

Пример 36. Вернемся к примеру 27. Определим наличие автокорреляции методом рядов.

e_i	8,3	4,26	–12,46	–1,86	–7,38	5,26	–9,66	–2,26	8,34	7,46
Знак	+	+	–	–	–	+	–	–	+	+

Последовательность знаков указана во второй строке. У нас $n_1 = 5$ (5 плюсов), $n_2 =$ (5 минусов), $k = 5$ (5 рядов). Из специальных таблиц Сведа-Эйзенхарта по n_1, n_2 и $\alpha = 0{,}05$ находим числа $k_1 = 2$ и $k_1 = 10$. Так как $k_1 < k < k_2$ $(2 < 5 < 10)$, то автокорреляция отсутствует.

Задача 36. $n_1 = 12$, $n_2 = 8$, $k = 3$, $k_1 = 6$ и $k_2 = 16$. Определить наличие автокорреляции методом рядов.

§ 6.2. КРИТЕРИЙ ДАРБИНА-УОТСОНА

Это наиболее известный способ обнаружения автокорреляции первого порядка. Пусть n — число наблюдений, m — число факторов модели, уровень значимости $\alpha = 0{,}05$. Для n, m, α по таблицам распределения Дарбина-Уотсона находим числа d_l и d_u.

Статистика Дарбина-Уотсона $DW = \dfrac{\sum\limits_{i=2}^{n}(e_i - e_{i-1})^2}{\sum\limits_{i=1}^{n} e_i^2}$.

Если $DW < d_l$, то это свидетельствует о положительной автокорреляции остатков. Если $DW > 4 - d_l$, то это свидетельствует об отрицательной автокорреляции остатков. При $d_u < DW < 4 - d_u$ гипотеза об отсутствии автокорреляции остатков принимается. Если $d_l < DW < d_u$ или $4 - d_u < DW < 4 - d_l$, то гипотеза об отсутствии автокорреляции остатков не может быть ни принята, ни отвергнута.

Ограничения при использовании критерия Дарбина-Уотсона:

1) $\beta_0 \neq 0$;

2) случайные отклонения определяются по авторегрессионной схеме первого порядка AR(1), то есть $\varepsilon_i = \rho\varepsilon_{i-1} + v_i$, где v_i — случайный член;

3) статистические данные должны иметь одинаковую периодичность (не должно быть пропусков в наблюдениях);

4) среди факторов не должно быть лаговых переменных (то есть переменных, влияние которых характеризуется определенным запаздыванием).

Пример 37. Вернемся к примеру 27. Определим наличие автокорреляции с помощью критерия Дарбина-Уотсона. $\sum_{i=1}^{n} e_i^2 = 551{,}52$.

Номер	e_i	$e_i - e_{i-1}$	$(e_i - e_{i-1})^2$
1	8,3		
2	4,26	−4,04	16,32
3	−12,46	−16,72	279,56
4	−1,86	10,6	112,36
5	−7,38	−5,52	30,47
6	5,26	12,64	159,77
7	−9,66	−14,92	222,61
8	−2,26	7,4	54,76
9	8,34	10,6	112,36
10	7,46	−0,88	0,77
Сумма			988,98

Заполняем таблицу. Из каждого числа 2-го столбца вычитаем предыдущее число 2-го столбца и результат пишем в 3-м столбце. В 4-м столбце числа округляем до двух знаков после запятой. Статистика Дарбина-Уотсона:

$$DW = \frac{\sum_{i=2}^{2}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{988{,}98}{551{,}52} \approx 1{,}793.$$

По таблице распределения Дарбина-Уотсона находим $d_l = 0{,}697$ и $d_u = 1{,}641$. Тогда $4 - d_u = 4 - 1{,}641 = 2{,}359$.

Так как $d_u < DW < 4 - d_u$ $(1{,}641 < 1{,}793 < 2{,}359)$, то гипотеза об отсутствии автокорреляции остатков не отклоняется на уровне значимости 0,05. Это является одним из подтверждений высокого качества модели.

Задача 37. В задаче 27 определить наличие автокорреляции с помощью критерия Дарбина-Уотсона. Уровень значимости 0,05.

§ 6.3. МЕТОДЫ УСТРАНЕНИЯ АВТОКОРРЕЛЯЦИИ

Возможно, автокорреляция вызвана отсутствием в модели важной объясняющей переменной. Нужно попытаться оп-

ределить данный фактор и включить его в модель. Также можно попробовать изменить форму зависимости. Но если все разумные процедуры изменения спецификации модели исчерпаны, а автокорреляция имеет место, то можно воспользоваться авторегрессионным преобразованием.

Для простоты ограничимся моделью парной линейной регрессии и авторегрессионной схемой первого порядка AR(1). $y = \beta_0 + \beta_1 x + \varepsilon$. Вместо переменных y, x рассмотрим переменные y^*, x^*, значения которых вычисляются по правилу $y_i^* = y_i - \rho y_{i-1}$, $x_i^* = x_i - \rho x_{i-1}$, $i = 2, \ldots, n$, $\rho \approx 1 - DW/2$.

Поправки Прайса-Винстена:

$$x_1^* = x_1\sqrt{1-\rho^2}, \quad y_1^* = y_1\sqrt{1-\rho^2}.$$

Положим $\beta_0^* = \beta_0(1-\rho)$. Тогда по таблице значений переменных y^*, x^* оцениваются коэффициенты уравнения $y^* = \beta_0^* + \beta_1 x^*$. Затем получаем $\beta_0 = \beta_0^*/(1-\rho)$.

Пример 38. Оцениваются коэффициенты уравнения $y = a + bx$, где значения переменных x, y — первые два столбца таблицы.

x_i	y_i	$x_i^* = x_i - 0,31x_{i-1}$	$y_i^* = y_i - 0,31y_{i-1}$
1,31	1,12	1,25	1,06
2,21	–0,36	1,80	–0,71
1,37	1,41	0,68	1,52
1,87	0,79	1,45	0,35
1,53	0,87	0,95	0,63
2,14	–0,11	1,67	–0,38
2,26	0,1	1,60	0,13
1,31	1,63	0,61	1,60
1,76	–0,07	1,35	–0,58
1,28	0,93	0,73	0,95
1,88	0,44	1,48	0,15
1,46	1,24	0,88	1,10
2,22	0,09	1,77	–0,29
1,75	0,77	1,06	0,74
1,29	1,64	0,75	1,40
1,99	0,54	1,59	0,03
2,27	–0,3	1,65	–0,47
1,29	1,43	0,59	1,52
2,28	–0,07	1,88	–0,51
1,84	0,58	1,13	0,60
2,05	0,22	1,48	0,04
2,17	0,11	1,53	0,04
1,98	0,25	1,31	0,22
1,28	2	0,67	1,92
1,29	1,67	0,89	1,05

На основании критерия Дарбина-Уотсона гипотеза об отсутствии автокорреляции остатков не может быть ни принята, ни отвергнута. Применяется авторегрессионная схема первого порядка AR(1).

$DW = 1{,}381$. $\rho \approx 1 - DW/2 = 1 - 1{,}381/2 \approx 0{,}31$.

$y_1^* = y_1\sqrt{1-\rho^2} = 1{,}12\sqrt{1-0{,}31^2} \approx 1{,}06$,

$x_1^* = x_1\sqrt{1-\rho^2} = 1{,}31\sqrt{1-0{,}31^2} \approx 1{,}25$.

Заполняем таблицу. Из каждого элемента 1-го столбца вычитаем предыдущее число 1-го столбца, умноженное на 0,31, и результат пишем в 3-м столбце (округляем до двух цифр после запятой). Аналогично для 2-го и 4-го столбцов.

По МНК находим коэффициенты уравнения $y^* = a^* + bx^*$. Тогда $a = a^*/(1-\rho) = a^*/(1-0{,}31) = a^*/0{,}69$.

Задача 38. Найти коэффициенты a, b в примере 38.

Глава 7

МУЛЬТИКОЛЛИНЕАРНОСТЬ

Мультиколлинеарность — это линейная взаимосвязь двух или нескольких объясняющих переменных. Каковы последствия мультиколлинеарности?

1) Большие стандартные ошибки S_{b_i}, что расширяет доверительные интервалы теоретических коэффициентов уравнения линейной регрессии.

2) Уменьшаются статистики $t_{b_i} = b_i/S_{b_i}$, поэтому возможен вывод о статистической незначимости коэффициента β_i.

3) b_i и S_{b_i} становятся очень чувствительными к малейшим изменениям данных.

4) Затрудняется определение вклада каждой из объясняющих переменных в объясняемую уравнением линейной регрессии дисперсию результативного признака.

5) Возможно получение неверного знака у коэффициентов регрессии.

§ 7.1. УСТАНОВЛЕНИЕ МУЛЬТИКОЛЛИНЕАРНОСТИ

Высокий коэффициент детерминации и низкие статистики t_{b_i} некоторых переменных свидетельствуют о наличии мультиколлинеарности. Также о наличии мультиколлинеарности свидетельствуют высокие значения частных коэффициентов корреляции (коэффициенты корреляции между двумя переменными, очищенные от влияния других переменных).

Например, при трех факторах x_1, x_2, x_3 частный коэффициент корреляции между x_1 и x_2 равен:

$$r_{12.3} = \frac{(r_{12} - r_{13} \times r_{23})}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}}.$$

Аналогично вычисляются $r_{13.2}$ и $r_{23.1}$.

Пример 39. В модели три фактора x_1, x_2, x_3. Коэффициенты корреляции $r_{12} = 0,44$, $r_{13} = -0,35$, $r_{23} = 0,51$. Найдем частный коэффициент корреляции между x_1 и x_2.

$$r_{12.3} = \frac{(r_{12} - r_{13} \times r_{23})}{\sqrt{(1 - r^2_{13})(1 - r^2_{23})}} = \frac{0,44 - (-0,35) \times 0,51}{\sqrt{(1 - (-0,35)^2)(1 - 0,51^2)}} \approx$$

$\approx 0,768.$

Задача 39. В модели три фактора x_1, x_2, x_3. Коэффициенты корреляции $r_{12} = 0,42$, $r_{13} = -0,36$, $r_{23} = 0,53$. Найти частный коэффициент корреляции между x_1 и x_2.

Замечание. Excel позволяет вычислить корреляционную матрицу (матрицу из попарных коэффициентов корреляции)

$$\begin{bmatrix} 1 & & & \\ r_{12} & 1 & & \\ \dots & \dots & \dots & \dots \\ r_{1m} & \dots r_{m-1,m} & 1 & \end{bmatrix}$$

для любого количества переменных.

Сервис → *Анализ данных* → *Корреляция* → *OK*. Появляется диалоговое окно, которое нужно заполнить. Рядом со словом *Группирование* нужно указать, как расположены исходные данные (*по строкам* или *по столбцам*). *OK*.

§ 7.2. МЕТОДЫ УСТРАНЕНИЯ МУЛЬТИКОЛЛИНЕАРНОСТИ

Иногда мультиколлинеарность не является таким уж серьезным «злом», чтобы прилагать существенные усилия по ее выявлению и устранению. Все зависит от целей исследования. Если основная задача модели — прогноз будущих значений результативного признака, то при $R^2 \geq 0,9$ наличие мультиколлинеарности обычно не сказывается на прогнозных качествах модели. Единого метода устранения мультиколлинеарности не существует. Простейшим методом устранения мультиколлинеарности является исключение из модели ряда коррелированных переменных. В прикладных моделях лучше не сокращать число факторов до тех пор, пока мультиколлинеарность не станет серьезной проблемой. Иногда для уменьшения мультиколлинеарности достаточно увеличить объем выборки. Но при этом может усилиться автокорреляция. Иногда проблема мультиколлинеарности может быть решена путем изменения спецификации модели.

Глава 8

ФИКТИВНЫЕ ПЕРЕМЕННЫЕ

Очень часто в регрессионных моделях в качестве объясняющих переменных используют не только количественные (определяются численно), но и качественные. Обычно в моделях влияние качественного фактора выражается в виде фиктивной переменной, которая отражает два противоположных состояния качественного фактора. Например,

$$D = \begin{cases} 0, & \text{если сотрудник не имеет высшего образования,} \\ 1, & \text{если сотрудник имеет высшее образование.} \end{cases}$$

Модели, в которых объясняющие переменные носят как количественный, так и качественный характер, называются ANCOVA-моделями (моделями ковариационного анализа). Если качественная переменная имеет k альтернативных значений, то при моделировании используют только $k-1$ фиктивную переменную. Значения фиктивной переменной можно менять на противоположные. Суть модели от этого не изменится.

Пример 40. Исследуется надежность станков трех производителей a, b, c. При этом учитывается возраст станка M (в месяцах) и время H (в часах) безаварийной работы до последней поломки. Выборка из 40 станков дала следующие результаты.

Фирма	H	M	F	R	Фирма	H	M	F	R
a	280	23	0	0	a	200	52	0	0
b	230	30	1	0	b	265	20	1	0
c	112	65	1	1	c	148	70	1	1
a	176	69	0	0	c	150	62	1	1
c	90	75	1	1	b	176	40	1	0
a	176	63	0	0	a	123	66	0	0
b	216	25	1	0	a	245	20	0	0
c	110	75	1	1	c	176	39	1	1
b	45	75	1	0	b	260	25	1	0

Фирма	H	M	F	R
a	236	48	0	0
a	205	59	0	0
a	240	25	0	0
b	65	69	1	0
a	115	71	0	0
c	200	26	1	1
b	126	45	1	0
a	225	40	0	0
c	210	30	1	1
b	45	69	1	0
a	260	30	0	0
b	220	22	1	0
b	194	33	1	0
c	156	48	1	1
a	100	75	0	0
b	240	21	1	0
a	170	56	0	0
c	116	58	1	1
b	120	40	1	0
a	240	37	0	0
b	88	56	1	0
a	120	67	0	0

У уравнения регрессии $H = \beta_0 + \beta_1 M + \varepsilon$ без учета различия станков различных фирм невысокий коэффициент детерминации $R^2 = 0{,}686$. Поэтому нужно учитывать производителя станков. Качественная переменная «Производитель станков» может принимать $k = 3$ значения (a, b, c). Поэтому нужно ввести в модель $k - 1 = 3 - 1 = 2$ фиктивных переменных F и R.

$$F = \begin{cases} 0, & \text{если производитель } a, \\ 1, & \text{если производитель } b \text{ или } c. \end{cases}$$

$$R = \begin{cases} 0, & \text{если производитель } a \text{ или } b, \\ 1, & \text{если производитель } c. \end{cases}$$

Для производителя a $F = R = 0$, для производителя b $F = 1$, $R = 0$, для производителя c $F = R = 1$.

Теперь нужно оценить коэффициенты уравнения $H = \beta_0 + \beta_1 M + \gamma_1 F + \gamma_2 R$ (см. результаты *главы 4*).

Задача 40. В примере 40 оценить коэффициенты уравнения $H = \beta_0 + \beta_1 M + \gamma_1 F + \gamma_2 R$.

Пусть рассматривается уравнение $y = \beta_0 + \beta_1 x$ и в модель решено ввести фиктивную переменную D.

Это можно сделать двумя способами: $y = \beta_0 + \beta_1 x + \gamma_1 D$ и $y = \beta_0 + \beta_1 x + \gamma_1 D + \gamma_2 Dx$.

Коэффициенты γ_1 и γ_2 называются *дифференциальным свободным членом* и *дифференциальным угловым коэффициентом* соответственно.

Фиктивная переменная D во втором уравнении используется как в аддитивном ($\gamma_1 D$), так и в мультипликативном виде ($\gamma_2 Dx$), что позволяет фактически разбивать рассматриваемую зависимость на две части, связанные с периодами изменения рассматриваемой в модели переменной.

Пример 41. Исследуется эффективность лекарств y в зависимости от x (возраст пациента). При этом сравнивается эффективность лекарств a и b.

Лек-во	y	x	D	Dx	Лек-во	y	x	D	Dx
a	54	69	0	0	b	30	40	1	40
b	30	48	1	48	b	23	41	1	41
a	58	73	0	0	a	21	55	0	0
b	66	64	1	64	b	43	45	1	45
b	67	60	1	60	a	38	58	0	0
a	64	62	0	0	b	43	58	1	58
a	67	70	0	0	a	43	64	0	0
a	33	52	0	0	b	45	55	1	55
a	33	63	0	0	b	48	57	1	57
b	42	48	1	48	a	48	63	0	0
b	33	46	1	46	a	53	60	0	0
a	28	55	0	0	b	58	62	1	62

Вводится фиктивная переменная D:

$$D = \begin{cases} 0, \text{ если лекарство } a, \\ 1, \text{ если лекарство } b. \end{cases}$$

Возможен один из трех вариантов: $y = \beta_0 + \beta_1 x$, $y = \beta_0 + \beta_1 x + \gamma_1 D$ или $y = \beta_0 + \beta_1 x + \gamma_1 D + \gamma_2 Dx$.

Задача 41. Какой из вариантов, по вашему мнению, предпочтительнее?

Глава 9

НЕЛИНЕЙНЫЕ СВЯЗИ

Многие экономические зависимости не являются линейными по своей сути, и поэтому их моделирование линейными уравнениями регрессии не приведет к положительному результату. Мы ограничимся нелинейными моделями, допускающими сведение к линейным моделям. Это *линейные относительно параметров модели*.

Пример 42. Дана зависимость $y = Ax^{\beta}$, где A и β — константы, подлежащие определению.

Тогда $\ln y = \ln(Ax^{\beta}) = \ln A + \ln x^{\beta} = \ln A + \beta \ln x$.

Положим $z = \ln y$, $\beta_0 = \ln A$, $t = \ln x$. Тогда $z = \beta_0 + \beta t$. Это линейное уравнение. Мы можем оценить коэффициенты β_0, β, а затем найти оценку для $A = e^{\beta_0}$.

Пример 43. Дана зависимость $y = AK^{\alpha}L^{\beta}$, где A, α и β — константы, подлежащие определению.

Тогда $\ln y = \ln(AK^{\alpha}L^{\beta}) = \ln A + \alpha \ln K + \beta \ln L$.

Положим $z = \ln y$, $\beta_0 = \ln A$, $x_1 = \ln K$, $x_2 = \ln L$. Тогда $z = \beta_0 + \alpha x_1 + \beta x_2$. Это линейная модель. Мы можем оценить коэффициенты β_0, α, β, а затем найти оценку для $A = e^{\beta_0}$.

Задача 42. $\ln y = \beta_0 + \beta_1 x$. Свести к линейной модели.

Задача 43. $y = \beta_0 + \beta_1 \ln x$. Свести к линейной модели.

Пример 44. $y = \beta_0 + \beta_1 x + \beta_2 x^2$.

Положим $x_1 = x$, $x_2 = x^2$. Получилась модель множественной линейной регрессии $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Задача 44. $y = \beta_0 + \beta_1 \times 1/x$. Свести к линейной модели.

Пример 45. Дана зависимость $y = AK^{\alpha}L^{\beta}e^{\gamma t}$, где A, α, β и γ — константы, подлежащие определению.

Тогда $\ln y = \ln(AK^{\alpha}L^{\beta}e^{\gamma t}) = \ln A + \alpha \ln K + \beta \ln L + \gamma t$.

Положим $z = \ln y$, $\beta_0 = \ln A$, $x_1 = \ln K$, $x_1 = \ln L$.

Тогда $z = \beta_0 + \alpha x_1 + \beta x_2 + \gamma t$. Это линейная модель.

Мы можем оценить коэффициенты β_0, α, β, γ, а затем найти оценку для $A = e^{\beta_0}$.

Задача 45. $y = \beta_0 e^{\beta x}$. Свести к линейной модели.

Глава 10

ПОРЯДКОВЫЕ ИСПЫТАНИЯ

До этого момента мы работали с данными, которые можно измерить. Но бывают ситуации, когда важнее упорядочивание данных, чем прямое измерение. Это — *порядковые испытания*, а сами данные называются порядковыми. Для порядковых данных составляются две последовательности одинаковой длины $n \geq 10$. Нас интересует, есть ли между ними связь. Задается доверительная вероятность p. $\alpha = 1 - p$.

H_0: между двумя последовательностями нет связи, они не согласованы друг с другом.

H_1: между двумя последовательностями существует некая связь.

По α находим граничную точку $z_\alpha > 0$. Для последовательностей вычисляем ранговый коэффициент корреляции Спирмена r_s. Статистика $z = r_s \sqrt{n-1}$.

Пример 46. Два человека дегустируют 10 сортов чая. Каждый из них расположил эти сорта в порядке убывания предпочтений (второй и третий столбцы). Есть ли какая-нибудь связь между этими результатами? Доверительная вероятность $p = 95\%$.

Сорт чая	Дегустатор 1	Дегустатор 2	d	d^2
А	6	5	1	1
Б	4	6	−2	4
В	3	4	−1	1
Г	10	7	3	9
Д	5	1	4	16
Е	1	2	−1	1
Ж	8	8	0	0
З	2	3	−1	1
И	7	9	−2	4
К	9	10	−1	1
Сумма	—	—	—	38

d — это разность между значениями дегустаторов для одного и того же сорта чая.

H_0: между результатами этих исследований нет связи, они не согласованы друг с другом.

H_1: между результатами исследований существует некая связь.

$p = 0{,}95$, $\alpha = 1 - p = 1 - 0{,}95 = 0{,}05 \Rightarrow z_\alpha = z_{0,05} = 1{,}645$. Ранговый коэффициент корреляции Спирмена:

$$r_s = 1 - 6\frac{\sum d^2}{n(n^2-1)} = 1 - 6\frac{38}{10(10^2-1)} \approx 0{,}77.$$

Статистика $z = r_s\sqrt{n-1} = 0{,}77\sqrt{10-1} = 2{,}31 > 1{,}645$.

Мы отвергаем гипотезу H_0 и принимаем гипотезу H_1 на уровне значимости 5%. Между результатами исследований существует некая связь.

Задача 46. Два человека дегустируют 10 сортов чая. Каждый из них расположил эти сорта в порядке убывания предпочтений (второй и третий столбцы). Есть ли какая-нибудь связь между этими результатами? Доверительная вероятность $p = 99\%$.

Сорт чая	Дегустатор 1	Дегустатор 2
А	1	2
Б	7	6
В	5	3
Г	6	7
Д	2	1
Е	3	4
Ж	4	5
З	9	10
И	8	8
К	10	9

Глава 11

ВРЕМЕННЫЕ РЯДЫ

В этой главе мы рассмотрим возможность использования данных за прошлые периоды для прогнозирования.

Множество данных, где время является независимой переменной, называется *временным рядом*. Будут рассмотрены аддитивные и мультипликативные модели.

Общее изменение со временем значений результативного признака называется *трендом*. Мы рассмотрим модели *линейного тренда*, то есть параметры тренда можно рассчитать с помощью модели линейной регрессии.

Сезонная вариация — это повторение данных через небольшой промежуток времени. Под «сезоном» можно понимать и день, и неделю, и месяц, и квартал. Если же промежуток времени будет длительным, то это — *циклическая вариация*. Мы остановимся на изучении данных для небольших интервалов времени, поэтому циклическую вариацию исключим из рассмотрения.

Сначала на основании прошлых данных определяется сезонная вариация. Исключив сезонную вариацию (проведя так называемую *десезонализацию данных*), с помощью модели линейной регрессии находим уравнение тренда. По уравнению тренда и прошлым данным вычисляем величины ошибок. Это среднее абсолютное отклонение $MAD = \sum|e_t|/n$ и среднеквадратическая ошибка $MSE = \sum e_t^2/n$, где e_t — это разность фактического и прогнозного значений в момент времени t, n — число наблюдений.

§ 11.1. АНАЛИЗ АДДИТИВНОЙ МОДЕЛИ

Для аддитивной модели фактическое значение A = трендовое значение T + сезонная вариация S + ошибка E.

Пример 47. В таблице указан объем продаж (тыс. руб.) за последние 11 кварталов. Дадим на основании этих данных прогноз объема продаж на следующие два квартала.

Квартал	1	2	3	4	5	6	7	8	9	10	11
Объем продаж	4	6	4	5	10	8	7	9	12	14	15

На первом шаге нужно исключить влияние сезонной вариации. Воспользуемся методом скользящей средней. Заполним таблицу.

Номер квартала	Объем продаж	Скользящая средняя за 4 квартала	Центрированная скользящая средняя	Оценка сезонной вариации
1	4			
2	6			
3	4	4,75	5,5	−1,5
4	5	6,25	6,5	−1,5
5	10	6,75	7,125	2,875
6	8	7,5	8	0
7	7	8,5	8,75	−1,75
8	9	9	9,75	−0,75
9	12	10,5	11,5	0,5
10	14	12,5		
11	15			

1 год = 4 квартала. Поэтому найдем среднее значение объема продаж за 4 последовательных квартала.

Для этого нужно сложить 4 последовательных числа из 2-го столбца, эту сумму разделить на 4 (количество слагаемых) и результат записать в 3-й столбец напротив третьего слагаемого.

$(4 + 6 + 4 + 5)/4 = 4{,}75$ (пишем напротив 4).

$(6 + 4 + 5 + 10)/4 = 6{,}25$ (пишем напротив 5). И т. д.

Полусумму двух соседних чисел из 3-го столбца запишем в 4-й столбец напротив верхнего из них. Если при заполнении 3-го столбца скользящая средняя вычислялась для нечетного числа сезонов, то результат записывается напротив среднего слагаемого и данные не надо центрировать (то есть не надо заполнять 4-й столбец). 5-й столбец — это разность 2-го и 4-го столбцов (2-го и 3-го столбцов, если скользящая средняя вычислялась для нечетного числа сезонов).

Заполним следующую таблицу. Оценки сезонной вариации запишем под соответствующим номером квартала в году. В каждом столбце вычисляем среднее = (сумма чисел в столбце)/(количество чисел в столбце). Результат пишем в строке «Среднее» (округления до одной цифры после запятой). Сумма чисел в строке «Среднее» равна −1.

Скорректируем значения в строке «Среднее», чтобы общая сумма была равна 0. Это необходимо, чтобы усреднить значения сезонной вариации в целом за год. Корректирующий фактор вычисляется следующим образом: сумма оценок сезонных вариаций (−1) делится на число кварталов в году (4). Поэтому из каждого числа этой строки надо вычесть −1/4 = −0,25. Так как у нас округления до одной цифры после запятой, то из нечетных столбцов вычтем −0,3, а из четных столбцов вычтем −0,2. В последней строке получены значения сезонной вариации для соответствующего квартала года.

	Номер квартала в году				
	1	2	3	4	
			−1,5	−1,5	
	2,875	0	−1,75	−0,75	
	0,5				Сумма
Среднее	1,7	0,0	−1,6	−1,1	−1
Скорректированная сезонная вариация	2,0	0,2	−1,3	−0,9	0,0

Исключим сезонную вариацию из фактических данных. Проведем десезонализацию данных.

Номер квартала	Объем продаж A	Сезонная вариация S	Десезонализированный объем продаж $A-S=T+E$
1	4	2	2
2	6	0,2	5,8
3	4	−1,3	5,3
4	5	−0,9	5,9
5	10	2	8
6	8	0,2	7,8
7	7	−1,3	8,3
8	9	−0,9	9,9
9	12	2	10
10	14	0,2	13,8
11	15	−1,3	16,3

Из чисел 2-го столбца вычитаем числа 3-го столбца и результат пишем в 4-м столбце.

Уравнение линии тренда $T = a + bx$.

Используя результаты *главы 3*, найдем коэффициенты a и b по данным первого и последнего столбцов.

Трендовое значение объема продаж = 1,9 + 1,1 × (номер квартала).

Теперь займемся расчетом ошибок.

Из чисел 3-го столбца вычитаем числа 4-го столбца и результат пишем в 5-м столбце. Среднее абсолютное отклонение $MAD = \sum|e_t|/n = 11{,}6/11 \approx 1{,}1$, среднеквадратическая ошибка $MSE = \sum e_t^2/n = 16{,}58/11 \approx 1{,}5$. Мы видим, что ошибки достаточно велики. Это скажется на качестве прогноза.

Номер квартала	Объем продаж A	Десезонализированный объем продаж $A-S=T+E$	Трендовое значение	Ошибка e_t	$\|e_t\|$	e_t^2
1	4	2	3	−1	1	1
2	6	5,8	4,1	1,7	1,7	2,89
3	4	5,3	5,2	0,1	0,1	0,01
4	5	5,9	6,3	−0,4	0,4	0,16
5	10	8	7,4	0,6	0,6	0,36
6	8	7,8	8,5	−0,7	0,7	0,49
7	7	8,3	9,6	−1,3	1,3	1,69
8	9	9,9	10,7	−0,8	0,8	0,64
9	12	10	11,8	−1,8	1,8	3,24
10	14	13,8	12,9	0,9	0,9	0,81
11	15	16,3	14	2,3	2,3	5,29
				Сумма	11,6	16,58

Дадим прогноз объема продаж на следующие два квартала. Мы считаем, что тенденция, выявленная по прошлым данным, сохранится и в ближайшем будущем. Подставляем номера кварталов в формулу и учитываем сезонную вариацию.

Прогноз объема продаж в 12-м квартале:
$(1{,}9 + 1{,}1 \times 12) + (-0{,}9) = 14{,}2$ тыс. руб.
Прогноз объема продаж в 13-м квартале:
$(1{,}9 + 1{,}1 \times 13) + 2 = 18{,}2$ тыс. руб.

Задача 47. В таблице указан объем продаж (тыс. руб.) за последние 11 кварталов. Дать на основании этих данных прогноз объема продаж на следующие два квартала.

Квартал	1	2	3	4	5	6	7	8	9	10	11
Объем продаж	4	5	5	6	9	9	8	10	11	13	16

§ 11.2. АНАЛИЗ МУЛЬТИПЛИКАТИВНОЙ МОДЕЛИ

В некоторых временных рядах значение сезонной вариации — это определенная доля трендового значения, то есть сезонная вариация увеличивается с возрастанием значений

тренда. В таких случаях используется мультипликативная модель. Для мультипликативной модели фактическое значение A = трендовое значение $T \times$ сезонная вариация $S \times$ ошибка E.

Пример 48. В таблице указан объем продаж (тыс. руб.) за последние 11 кварталов. Дадим на основании этих данных прогноз объема продаж на следующие два квартала.

Квартал	1	2	3	4	5	6	7	8	9	10	11
Объем продаж	63	74	79	120	67	79	88	130	69	82	90

Числа 2-го столбца приведенной далее таблицы делим на числа 4-го столбца и результат (округляем до трех цифр после запятой) запишем в 5-й столбец.

Номер квартала	Объем продаж	Скользящая средняя за 4 квартала	Центрированная скользящая средняя	Оценка сезонной вариации
1	63			
2	74			
3	79	84	84,5	0,935
4	120	85	85,625	1,401
5	67	86,25	87,375	0,767
6	79	88,5	89,75	0,880
7	88	91	91,25	0,964
8	130	91,5	91,875	1,415
9	69	92,25	92,5	0,746
10	82	92,75		
11	90			

	Номер квартала в году				
	1	2	3	4	
			0,935	1,401	
	0,767	0,880	0,964	1,415	
	0,746				Сумма
Среднее	0,756	0,880	0,950	1,408	3,994
Скорректированная сезонная вариация	0,757	0,881	0,952	1,410	4,000

Значения сезонной вариации — это доли, число сезонов равно 4. Поэтому необходимо, чтобы сумма средних была равна 4. У нас же получилось 3,994. Следовательно, итоговые коэффициенты сезонности нужно умножить на множитель 4/3,994. В последней строке указаны окончательные коэффициенты сезонности.

Как показывают полученные оценки, в 1-м, 2-м и 3-м кварталах года объем продаж снижается соответственно на 24,3%, 11,9% и 4,8% от соответствующих трендовых значений. В 4-м квартале года объем продаж увеличивается на 41% от соответствующего трендового значения.

Исключим сезонную вариацию из фактических данных. Проведем десезонализацию данных. Числа 2-го столбца делим на числа 3-го столбца, результат округляем до одной цифры после запятой и пишем в 4-й столбец.

Номер квартала	Объем продаж A	Коэффициент сезонности S	Десезонализированный объем продаж $A/S=T\times E$
1	63	0,757	83,2
2	74	0,881	84,0
3	79	0,952	83,0
4	120	1,41	85,1
5	67	0,757	88,5
6	79	0,881	89,7
7	88	0,952	92,4
8	130	1,41	92,2
9	69	0,757	91,1
10	82	0,881	93,1
11	90	0,952	94,5

Уравнение линии тренда $T = a + bx$.

Используя результаты *главы 3*, найдем коэффициенты a и b по данным первого и последнего столбцов.

Трендовое значение объема продаж $= 81{,}6 + 1{,}2 \times$ (номер квартала).

Теперь займемся расчетом ошибок.

Номер квартала	Объем продаж A	Коэффициент сезонности S	Десезонализированный объем продаж $A/S=T\times E$	Трендовое значение	Ошибка e_t	$\lvert e_t \rvert$	e_t^2
1	63	0,757	83,2	82,8	0,4	0,4	0,16
2	74	0,881	84,0	84	0,0	0,0	0,00
3	79	0,952	83,0	85,2	−2,2	2,2	4,84
4	120	1,41	85,1	86,4	−1,3	1,3	1,69
5	67	0,757	88,5	87,6	0,9	0,9	0,81
6	79	0,881	89,7	88,8	0,9	0,9	0,81
7	88	0,952	92,4	90	2,4	2,4	5,76
8	130	1,41	92,2	91,2	1,0	1,0	1,00
9	69	0,757	91,1	92,4	−1,3	1,3	1,69
10	82	0,881	93,1	93,6	−0,5	0,5	0,25
11	90	0,952	94,5	94,8	−0,3	0,3	0,09
					Сумма	11,2	17,10

Среднее абсолютное отклонение $MAD = \sum|e_t|/n = 11{,}1/11 \approx 1$, среднеквадратическая ошибка $MSE = \sum e_t^2/n = 17{,}10/11 \approx 1{,}6$. Мы видим, что ошибки малы и составляют порядка 1%. Это позволяет получить хорошие краткосрочные прогнозы.

Дадим прогноз объема продаж на следующие два квартала. Мы считаем, что тенденция, выявленная по прошлым данным, сохранится и в ближайшем будущем. Подставляем номера кварталов в формулу и учитываем сезонную вариацию.

Прогноз объема продаж в 12-м квартале:
$(81{,}6 + 1{,}2 \times 12) \times 1{,}41 \approx 135{,}4$ тыс. руб.
Прогноз объема продаж в 13-м квартале:
$(81{,}6 + 1{,}2 \times 13) \times 0{,}757 \approx 73{,}6$ тыс. руб.

Задача 48. В таблице указан объем продаж (тыс. руб.) за последние 11 кварталов. Дать на основании этих данных прогноз объема продаж на следующие два квартала.

Квартал	1	2	3	4	5	6	7	8	9	10	11
Объем продаж	64	75	81	110	66	77	91	120	68	78	92

Замечание. Excel позволяет быстро вычислить оценки методом скользящей средней. *Сервис → Анализ данных → Скользящее среднее → ОК.* Появляется диалоговое окно, которое нужно заполнить. В графе *Интервал* вводится количество сезонов, для которых вычисляется скользящее среднее (по умолчанию это 3). Если требуется график, на котором будут указаны прогнозные и фактические значения, то нужно поставить «галочку» рядом со словосочетанием *Вывод графика. ОК.* Появляется итоговая таблица, которая содержит исходные данные и оценки, полученные методом скользящей средней. Если оценка находилась как среднее k слагаемых, то в таблице оценок она находится напротив последнего из этих k слагаемых.

Глава 12

ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ

При анализе временных рядов использовался метод скользящей средней, где все данные (и поздние, и ранние) были равноправны. Более правильным представляется способ, в котором данным приписываются веса: более поздним данным придается больший вес, чем более ранним. Этот метод обеспечивает быстрое получение прогноза на один период вперед и автоматически корректирует любой прогноз в свете различий между фактическим и спрогнозированным результатом.

§ 12.1. ПРОСТАЯ МОДЕЛЬ ЭКСПОНЕНЦИАЛЬНОГО СГЛАЖИВАНИЯ

Новый прогноз = $\alpha \times$(фактический результат в последний период) + $(1-\alpha)\times$(прогноз в последний период), то есть $F_{t+1} = \alpha A_t + (1-\alpha)F_t$. Константу сглаживания α исследователь выбирает из отрезка [0, 1]. В условиях стабильности часто $\alpha \in [0{,}2;\ 0{,}4]$.

Пример 49. Вернемся к примеру 47. Пусть $\alpha = 0{,}8$. Тогда $1 - \alpha = 1 - 0{,}8 = 0{,}2$. Предположим, что на первый квартал был дан прогноз 3. Дадим прогноз на 12-й квартал.

Заполним таблицу.

$F_{t+1} = \alpha A_t + (1-\alpha) = 0{,}8A_t + 0{,}2F_t$, то есть числа в каждой строке умножаем соответственно на 0,8 и 0,2, и результат пишем в следующей строке во втором столбце.

$0{,}8\times 4 + 0{,}2\times 3 = 3{,}8$.

$0{,}8\times 6 + 0{,}2\times 3{,}8 \approx 5{,}6$. И т. д.

Результат округляем до одной цифры после запятой.

A_t (фактически)	F_t (прогноз)
4	3
6	3,8
4	5,6
5	4,3
10	4,9
8	9
7	8,2
9	7,2
12	8,6
14	11,3
15	13,5
	14,7

Прогноз объема продаж на 12-й квартал — 14,7 тыс. руб.

Задача 49. В задаче 47 дать прогноз объема продаж на 12-й квартал методом простого экспоненциального сглаживания. $\alpha = 0,8$. $F_1 = 3$.

Замечание. Excel позволяет быстро провести простое экспоненциальное сглаживание. *Сервис* → *Анализ данных* → *Экспоненциальное сглаживание* → *ОК*. Появляется диалоговое окно, которое нужно заполнить. В графе *Фактор затухания* указать значение α (по умолчанию 0,3). *ОК*.

§ 12.2. ЭКСПОНЕНЦИАЛЬНОЕ СГЛАЖИВАНИЕ С ПОПРАВКОЙ НА ТРЕНД

Даем прогноз методом простого экспоненциального сглаживания, а затем корректируем его с учетом тренда по следующей формуле:

прогноз с учетом тренда FIT_t = прогноз F_t + тренд T_t.

Тренд $T_t = (1 - b)T_{t-1} + b(F_t - F_{t-1})$, где T_t и T_{t-1} — сглаженный тренд в периоды t и $t-1$ соответственно, b — выбранная константа сглаживания.

Начальное значение тренда может быть получено на основе предположения.

Пример 50. В примере 49 дадим прогноз объема продаж на 12-й квартал методом экспоненциального сглаживания с поправкой на тренд. Возьмем $b = 0,4$, $T_1 = 0$.

F_t	F_t-F_{t-1}	T_t	$FIT_t=F_t+T_t$
3	—	0	3
3,8	0,8	0,3	4,1
5,6	1,8	0,9	6,5
4,3	−1,3	0,0	4,3
4,9	0,6	0,3	5,2
9	4,1	1,8	10,8
8,2	−0,8	0,8	9,0
7,2	−1	0,1	7,3
8,6	1,4	0,6	9,2
11,3	2,7	1,4	12,7
13,5	2,2	1,7	15,2
14,7	1,2	1,5	16,2

Заполним таблицу. Из каждого числа 1-го столбца вычитаем предыдущее число 1-го столбца и результат запишем во 2-й столбец. Каждое число 3-го столбца есть сумма числа, умноженного на $1-b=1-0{,}4=0{,}6$, из предыдущей строки 3-го столбца и числа, умноженного на $b=0{,}4$, из этой же строки 2-го столбца. Результат округляем до одной цифры после запятой.

Прогноз объема продаж на 12-й квартал — 16,2 тыс. руб.

Задача 50. В задаче 49 дать прогноз объема продаж на 12-й квартал методом экспоненциального сглаживания с поправкой на тренд. $b = 0{,}4$, $T_1 = 0$.

Глава 13

СИСТЕМЫ ОДНОВРЕМЕННЫХ УРАВНЕНИЙ

Ряд экономических процессов моделируется несколькими уравнениями, содержащими как повторяющиеся, так и собственные переменные. Необходимо использовать системы одновременных уравнений.

Пример 51. Модель «спрос-предложение» содержит функции спроса, предложения и условие равновесия.

$$\begin{cases} q_t^D = \alpha_0 + \alpha_1 p_t + \varepsilon_{t1},\ \alpha_1 < 0 \\ q_t^S = \beta_0 + \beta_1 p_t + \varepsilon_{t2},\ \beta_1 > 0 \\ q_t^D = q_t^S. \end{cases}$$

Здесь первое уравнение — функция спроса, второе уравнение — функция предложения, третье уравнение — условие равновесия, p_t — цена товара в момент времени t, ε_{t1} и ε_{t2} — случайные составляющие.

Пример 52. Кейнсианская модель формирования доходов. Рассматривается закрытая экономика без государственных расходов.

$$\begin{cases} c_t = \beta_0 + \beta_1 y_t + \varepsilon_t, \\ y_t = c_t + i_t. \end{cases}$$

Здесь первое уравнение — функция потребления, второе уравнение — макроэкономическое тождество, y_t и i_t — значения совокупного выпуска и инвестиций в момент времени t соответственно, ε_t — случайная составляющая.

§ 13.1. СОСТАВЛЯЮЩИЕ СИСТЕМ ОДНОВРЕМЕННЫХ УРАВНЕНИЙ

Переменные в системах одновременных уравнений бывают двух видов: *эндогенные* (их значения определяются внутри модели) и *экзогенные* (внешние по отношению к модели, их значения считаются фиксированными).

Пример 53. В примере 52 c_t и y_t оцениваются внутри модели (эндогенные переменные), i_t задается вне модели (экзогенная переменная). Значения переменной i_t используются как заранее заданные. Из модели нельзя понять, как получаются значения переменной i_t.

Уравнения, составляющие исходную модель, называются *структурными уравнениями модели*. К ним относятся *поведенческие уравнения* (описывают взаимодействие между переменными) и *уравнения-тождества* (должны выполняться во всех случаях, не содержат параметров и случайных составляющих).

В приведенных уравнениях эндогенные переменные выражены через экзогенные и предопределенные (лаговые эндогенные переменные, значения которых определяются до рассмотрения соотношения).

Пример 54. Рассмотрим кейнсианскую модель формирования доходов.

$$\begin{cases} c_t = \beta_0 + \beta_1 y_t + \varepsilon_t, \\ y_t = c_t + i_t. \end{cases}$$

Это структурные уравнения. Эндогенные переменные: c_t и y_t. Экзогенная переменная i_t.

Подставим значение переменной c_t из первого уравнения во второе уравнение: $y_t = \beta_0 + \beta_1 y_t + i_t + \varepsilon_t$. Отсюда $y_t = \beta_0/(1-\beta_1) + i_t/(1-\beta_1) + \varepsilon_t/(1-\beta_1)$.

Тогда $c_t = \beta_0 + \beta_1 y_t + \varepsilon_t = \beta_0 + \beta_1(\beta_0/(1-\beta_1) + i_t/(1-\beta_1) + \varepsilon_t/(1-\beta_1)) + \varepsilon_t = \beta_0/(1-\beta_1) + \beta_1 i_t/(1-\beta_1) + \varepsilon_t/(1-\beta_1)$.

Приведенные уравнения:

$$\begin{cases} y_t = \beta_0/(1-\beta_1) + i_t/(1-\beta_1) + \varepsilon_t/(1-\beta_1), \\ c_t = \beta_0/(1-\beta_1) + \beta_1 i_t/(1-\beta_1) + \varepsilon_t/(1-\beta_1). \end{cases}$$

В левой части уравнений — только эндогенные переменные, в правой части — только экзогенная переменная и случайные отклонения.

§ 13.2. КОСВЕННЫЙ МЕТОД НАИМЕНЬШИХ КВАДРАТОВ (КМНК)

Непосредственное использование МНК для оценки параметров каждого из уравнений приводит к плохим результатам. Поэтому применяют другие методы. Например, косвенный метод наименьших квадратов.

По структурным уравнениям строят приведенные уравнения. Для приведенных уравнений по МНК находят оценки параметров и на их основании оценивают параметры структурных уравнений.

Пример 55. Рассмотрим модель «спрос-предложение» следующего вида:

$$\begin{cases} q_t^D = \alpha_0 + \alpha_1 p_t + \varepsilon_t, \\ q_t^S = \beta_0 + \beta_1 p_t + \beta_2 w_t + v_t, \\ q_t^D = q_t^S. \end{cases}$$

Здесь первое уравнение — функция спроса, второе уравнение — функция предложения, третье уравнение — условие равновесия, p_t и w_t — цена товара и зарплата в момент времени t соответственно, ε_t и v_t — случайные составляющие.

Имеются следующие результаты наблюдений.

p	10	15	5	8	4
q	6	6	18	12	8
w	2	6	2	7	4

Найдем оценки параметров этой системы уравнений с помощью КМНК.

Обозначим $q_t^S = q_t^S$ через q_t. Тогда

$$\begin{cases} q_t = \alpha_0 + \alpha_1 p_t + \varepsilon_t, \\ q_t = \beta_0 + \beta_1 p_t + \beta_2 w_t + v_t. \end{cases}$$

Приравняем правые части этих уравнений:

$$\alpha_0 + \alpha_1 p_t + \varepsilon_t = \beta_0 + \beta_1 p_t + \beta_2 w_t + v_t.$$

Отсюда:

$$p_t = (\beta_0 - \alpha_0)/(\alpha_1 - \beta_1) + \beta_2 w_t/(\alpha_1 - \beta_1) + (v_t - \varepsilon_t)/(\alpha_1 - \beta_1).$$

Введем обозначения:

$$\pi_{10} = (\beta_0 - \alpha_0)/(\alpha_1 - \beta_1),\ \pi_{11} = \beta_2/(\alpha_1 - \beta_1),$$
$$\psi_t = (v_t - \varepsilon_t)/(\alpha_1 - \beta_1).$$

Тогда $p_t = \pi_{10} + \pi_{11} w_t + \psi_t$.

Подставим это выражение для p_t в первое уравнение, раскроем скобки и введем следующие обозначения:

$$\pi_{20} = \alpha_0 + \alpha_1 \pi_{10},\ \pi_{21} = \alpha_1 \pi_{11},\ \phi_t = \alpha_1 \psi_t + \varepsilon_t.$$

Тогда $q_t = \pi_{20} + \pi_{21} w_t + \phi_t$.

Получаем систему из приведенных уравнений

$$\begin{cases} p_t = \pi_{10} + \pi_{11} w_t + \psi_t, \\ q_t = \pi_{20} + \pi_{21} w_t + \phi_t. \end{cases}$$

Оценим параметры каждого из этих уравнений при помощи МНК (*глава 3*).

$\pi_{11} = 0{,}75$, $\pi_{10} = 5{,}25$, $\pi_{21} = -0{,}46$, $\pi_{20} = 11{,}53$.

Тогда $\alpha_1 = \pi_{21}/\pi_{11} = -0{,}46/0{,}75 \approx -0{,}61$, $\alpha_0 = \pi_{20} - \alpha_1\pi_0 = 11{,}53 - (-0{,}61) \times 5{,}25 \approx 14{,}73$.

Мы не можем оценить коэффициенты β_i на основании полученных результатов. Возникает так называемая *проблема идентификации*.

Задача 51. Модель «спрос-предложение» содержит функции спроса, предложения

$$\begin{cases} q_t = \alpha_0 + \alpha_1 p_t + \alpha_2 y_t + \varepsilon_{t1}, \\ q_t = \beta_0 + \beta_1 p_t + \varepsilon_{t2}. \end{cases}$$

Здесь первое уравнение — функция спроса, второе уравнение — функция предложения, q_t, p_t, и y_t — количество товара, цена товара и доход потребителей в момент времени t соответственно, ε_{t1} и ε_{t2} — случайные составляющие. Эндогенные переменные: q_t, p_t. Экзогенная переменная y_t.

Имеются следующие результаты наблюдений.

p	1	2	3	4	5
q	8	10	7	5	1
w	2	4	3	5	2

Найти оценки параметров этой системы уравнений с помощью КМНК.

§ 13.3. ПРОБЛЕМА ИДЕНТИФИКАЦИИ

Проблема идентификации — это возможность численной оценки параметров структурных уравнений по оценкам коэффициентов приведенных уравнений. Можно ли найти оценки коэффициентов структурных уравнений по оценкам коэффициентов приведенных уравнений? Если да, то будут ли найденные оценки единственными?

Исходная система уравнений называется:

а) *идентифицируемой*, если по оценкам коэффициентов приведенных уравнений можно однозначно определить коэффициенты структурных уравнений;

б) *неидентифицируемой*, если по оценкам коэффициентов приведенных уравнений можно получить несколько вариантов значений коэффициентов структурных уравнений;

в) *сверхидентифицируемой* (переопределенной), если по оценкам коэффициентов приведенных уравнений невозможно определить коэффициенты структурных уравнений.

Пример 56. Система уравнений из примера 55 — сверхидентифицируемая система.

Пример 57. Система уравнений из задачи 51 — идентифицируемая система.

§ 13.4. НЕОБХОДИМЫЕ УСЛОВИЯ ИДЕНТИФИЦИРУЕМОСТИ

Пусть система одновременных уравнений включает в себя N уравнений относительно N эндогенных переменных. Система содержит M экзогенных либо предопределенных переменных. Пусть n и m — количество соответственно эндогенных и экзогенных переменных в проверяемом на идентифицируемость уравнении.

Первое необходимое условие идентифицируемости. Уравнение идентифицируемо, если $(N-n)+(M-m)\geq N-1$.

Второе необходимое условие идентифицируемости. Уравнение идентифицируемо, если $M-m\geq n-1$.

Пример 58. Рассмотрим модель денежного рынка

$$\begin{cases} r_t=\beta_0+\beta_1 y_t+\beta_2 m_t+\varepsilon_t, \\ y_t=\alpha_0+\alpha_1 r_t+v_t. \end{cases}$$

Здесь r_t, y_t, m_t — процентная ставка, ВВП и денежная масса в году t соответственно, ε_t и v_t — случайные составляющие. Эндогенные переменные: r_t и y_t ($N=2$). Экзогенная переменная m_t ($M=1$).

Первое уравнение содержит эндогенные переменные r_t и y_t ($n=2$) и экзогенную переменную m_t ($m=1$).

Проверим, имеет ли место $(N-n)+(M-m)\geq N-1$: $(2-2)+(1-1)\geq 2-1$ и $0\geq 1$ (ложно). Первое необходимое условие идентифицируемости не выполняется. Поэтому первое уравнение неидентифицируемо.

Второе уравнение содержит эндогенные переменные r_t и y_t ($n=2$) и не содержит экзогенной переменной m_t ($m=0$).

Проверим, имеет ли место $(N-n)+(M-m)\geq N-1$: $(2-2)+(1-0)\geq 2-1$ и $1=1$ (верно). Первое необходимое условие идентифицируемости выполняется. Поэтому второе уравнение точно идентифицируемо.

Пример 59. Рассмотрим модифицированную модель Кейнса

$$\begin{cases} c_t = \beta_0 + \beta_1 y_t + \varepsilon_t, \\ i_t = \gamma_0 + \gamma_1 y_t + \gamma_2 y_{t-1} + v_t, \\ y_t = c_t + i_t + g_t. \end{cases}$$

Здесь g_t — объем государственных расходов. Эндогенные переменные: c_t, i_t и y_t ($N = 3$). Экзогенные переменные: g_t и y_{t-1} ($M = 2$).

Первое уравнение эндогенные переменные c_t и y_t ($n = 2$) и не содержит экзогенных переменных ($m = 0$).

$(N-n)+(M-m) \geq N-1$, то есть $(3-2)+(2-0) \geq 3-1$ и $3 > 2$. Поэтому первое уравнение переопределено.

Второе уравнение содержит эндогенные переменные i_t и y_t ($n = 2$) и экзогенную переменную y_{t-1} ($m = 1$).

$(N-n)+(M-m) \geq N-1$, то есть $(3-2)+(2-1) \geq 3-1$ и $2 = 2$ (верно). Поэтому второе уравнение точно идентифицируемо.

Третье уравнение содержит эндогенные переменные c_t, i_t и y_t ($n = 3$) и экзогенную переменную g_t ($m = 1$).

$(N-n)+(M-m) \geq N-1$, то есть $(3-3)+(2-1) \geq 3-1$ и $1 \geq 2$ (ложно). Поэтому третье уравнение неидентифицируемо.

Задача 52. Рассматривается модифицированная модель «доход-потребление»

$$\begin{cases} c_t = \beta_0 + \beta_1 y_t + \beta_2 c_{t-1} + \varepsilon_t, \\ i_t = \gamma_0 + \gamma_1 r_t + v_t, \\ y_t = c_t + i_t + g_t. \end{cases}$$

Указать эндогенные и экзогенные переменные, определить идентифицируемость структурных уравнений, составить приведенную систему.

§ 13.5. ДВУХШАГОВЫЙ МНК (ДМНК)

Этот метод применяется к переопределенным уравнениям.

Пример 60. Вернемся к примеру 59.

Первое уравнение исходной системы $c_t = \beta_0 + \beta_1 y_t + \varepsilon_t$ переопределено, то есть по оценкам коэффициентов приведенных уравнений невозможно определить оценки коэффициентов β_0 и β_1. Для переменной y_t строим приведенное уравнение $y_t = \pi_{10} + \pi_{11} y_{t-1} + \pi_{12} g_t + w_t$ (w_t — случайное

отклонение), находим с помощью МНК оценки коэффициентов π_{10}, π_{11}, π_{12} и из уравнения получаем оценку $\hat{y}_t$ по экзогенным переменным y_{t-1} и g_t.

Из уравнения $c_t = \beta_0 + \beta_1 \hat{y}_t$ находим оценки коэффициентов β_0 и β_1 с помощью МНК.

Задача 53. Применить ДМНК в задаче 52.

Ответы

1. (77,78; 80,22).
2. 239.
3. (77,5; 80,5).
4. 240.
5. (0,09; 0,15).
6. 77860.
7. Автомат нужно отрегулировать.
8. Станок настроен правильно.
9. Выборка не противоречит утверждению производителя.
10. Выборка не противоречит утверждению производителя.
11. Риски инвестиций равны.
12. Автоматы фасуют чай в пачки разного среднего веса.
13. По первой технологии требуется в среднем больше времени для производства одной детали.
14. По первой технологии требуется в среднем больше времени для производства одной детали.
15. Побочные эффекты от нового лекарства у женщин возникают чаще, чем у мужчин.
16. Износоустойчивость шин одинакова.
17. Есть связь между оценками.
18. $y = 58{,}42 + 2{,}43x$.
19. 0,884; 0,781.

e_i	1,35	−1,95	4,92	−0,65	2,22	−0,38	−5,68	−3,22	−3,78	7,18

20. 71,79 тыс. руб.
21. Между переменными x, y есть линейная связь.
22. Между переменными x, y есть линейная связь.
23. (0,90; 3,96).
24. (67,04; 76,54).
25. (56,79; 86,79).
26. $y = 109{,}96 + 0{,}89x_1 - 11{,}14x_2$.
27. $S = 6{,}27$, $S_{b_0} = 10{,}61$, $S_{b_1} = 0{,}21$, $S_{b_2} = 1{,}26$.

28. (72,84; 147,08), (0,16; 1,62), (−15,54; −6,74).

29. Все коэффициенты статистически значимы.

30. 0,932. Коэффициент детерминации статистически значим.

31. Не ухудшилось.

32. Нужно строить единое уравнение регрессии.

33. Гетероскедастичность отсутствует.

34. Гетероскедастичность отсутствует.

35. $y = 0{,}40 + 0{,}01x$.

36. Автокорреляция отсутствует.

37. Гипотеза об отсутствии автокорреляции не может быть ни принята, ни отклонена.

38. $y = 3{,}77 - 1{,}72x$.

39. 0,772.

40. $y = 347{,}25 - 3{,}05M - 59{,}40F + 26{,}22R$.

41. $y = -53{,}19 + 1{,}58x + 14{,}84D$.

42. $z = \beta_0 + \beta_1 x,\ z = \ln y$.

43. $y = \beta_0 + \beta_1 z,\ z = \ln x$.

44. $y = \beta_0 + \beta_1 z,\ z = \ln 1/x$.

45. $z = A + \beta x,\ z = \ln y,\ \beta_0 = e^A$.

46. Между результатами исследований есть связь.

47. 14,3 тыс. руб., 17,7 тыс. руб.

48. 123 тыс. руб., 71,6 тыс. руб.

49. 15,3 тыс. руб.

50. 17,2 тыс. руб.

51. $q_t^s = -8{,}20 + 4{,}80p_t$. Функция спроса не идентифицируема.

52. Эндогенные переменные: c_t, i_t, y_t. Экзогенные переменные: c_{t-1}, r_t, g_t. Третье уравнение идентифицируемо, остальные — переопределены.

$$\begin{cases} c_t = \pi_{10} + \pi_{11} r_t + \pi_{12} g_t + \pi_{13} c_{t-1} + w_t, \\ i_t = \gamma_0 + \gamma_1 r_t + v_t, \\ y_t = \pi_{20} + \pi_{21} r_t + \pi_{22} g_t + \pi_{23} c_{t-1} + \delta_t. \end{cases}$$

где $\pi_{10} = (\beta_0 + \beta_1\gamma_0)/(1 - \beta_1)$, $\pi_{11} = \beta_1\gamma_1/(1 - \beta_1)$,
$\pi_{12} = \beta_1/(1 - \beta_1)$, $\pi_{13} = \beta_1/(1 - \beta_1) = \pi_{23}$,
$w_t = (\beta_1 v_t + \varepsilon_t)/(1 - \beta_1)$,
$\pi_{20} = \pi_{10} + \gamma_0$, $\pi_{21} = \pi_{11} + \gamma_1$, $\pi_{22} = \pi_{12} + 1$,
$\delta_t = w_t + v$.

53. Находим $\hat{c}_t$, $\hat{i}_t$ из приведенной системы и получаем $y_t = \hat{c}_t + \hat{i}_t + g_t$.

Программа учебного курса «Эконометрика»

1. Доверительный интервал для генеральной средней при известной генеральной дисперсии.
2. Объем выборки, необходимый для оценки генеральной средней при известной генеральной дисперсии.
3. Доверительный интервал для генеральной средней при неизвестной генеральной дисперсии.
4. Объем выборки, необходимый для оценки генеральной средней при неизвестной генеральной дисперсии.
5. Доверительный интервал для генеральной доли.
6. Объем выборки, необходимый для оценки генеральной доли.
7. Испытание гипотез, процедура испытания гипотез, односторонняя и двусторонняя проверки, статистика.
8. Испытание гипотезы на основе выборочной средней при известной генеральной дисперсии.
9. Испытание гипотезы на основе выборочной средней при неизвестной генеральной дисперсии.
10. Испытание гипотезы на основе выборочной доли.
11. Испытание гипотезы о двух генеральных дисперсиях, отношение дисперсий (F-критерий).
12. Сравнение средних величин двух выборок при известных генеральных дисперсиях.
13. Испытание гипотезы по выборочным средним (генеральные дисперсии неизвестны, случай равенства генеральных дисперсий).
14. Испытание гипотезы по выборочным средним (генеральные дисперсии неизвестны и не равны друг другу).
15. Испытание гипотезы по двум выборочным долям.
16. Испытание гипотез по спаренным данным (зависимые выборки).
17. Непараметрические испытания гипотез. Таблица сопряженности. Критерий Хи-квадрат. Поправка Йетса.
18. Простая модель линейной регрессии. Расчет коэффициентов в модели парной линейной регрессии.
19. Коэффициент корреляции Пирсона r. Объясненная, необъясненная и общая вариации переменной y. Коэффициент детерминации. Ошибки и остатки.
20. Предсказания и прогнозы на основе модели линейной регрессии.

21. Основные предпосылки в модели парной линейной регрессии.

22. Испытание гипотезы для оценки линейности связи на основе оценки коэффициента корреляции в генеральной совокупности.

23. Испытание гипотезы для оценки линейности связи на основе оценки показателя наклона линейной регрессии.

24. Доверительные интервалы в линейном регрессионном анализе. Доверительный интервал для показателя наклона линейной регрессии.

25. Доверительный интервал для среднего значения переменной y при заданном значении x.

26. Доверительный интервал для индивидуальных значений y при заданном значении x.

27. Множественная линейная регрессия. Основные предпосылки модели множественной линейной регрессии.

28. Расчет коэффициентов множественной линейной регрессии методом наименьших квадратов (МНК).

29. Стандартные ошибки коэффициентов в модели множественной линейной регрессии. Стандартная ошибка регрессии.

30. Интервальные оценки теоретического уравнения линейной регрессии.

31. Проверка статистической значимости коэффициентов уравнения линейной регрессии.

32. Проверка общего качества уравнения линейной регрессии. Коэффициент детерминации. Исправленный коэффициент детерминации.

33. Проверка равенства двух коэффициентов детерминации.

34. Проверка гипотезы о совпадении уравнений регрессии для двух выборок. Тест Чоу.

35. Регрессия и Excel.

36. Гетероскедастичность, ее последствия. Тест ранговой корреляции Спирмена.

37. Тест Голдфелда-Квандта.

38. Смягчение проблемы гетероскедастичности. Метод взвешенных наименьших квадратов (ВНК) в случае пропорциональности дисперсии отклонений квадрату независимой переменной.

39. Автокорреляция. Метод рядов. Таблица Сведа-Эйзенхарта.

40. Критерий Дарбина-Уотсона.

41. Устранение автокорреляции. Авторегрессионная схема первого порядка AR(1). Поправки Прайса-Винстена.

42. Мультиколлинеарность. Частный коэффициент корреляции. Корреляционная матрица. Методы устранения мультиколлинеарности.

43. Фиктивные переменные. ANCOVA-модели (модели ковариационного анализа). Дифференциальный свободный член и дифференциальный угловой коэффициент.

44. Нелинейные связи.

45. Порядковые испытания. Ранговый коэффициент корреляции Спирмена.

46. Элементы временного ряда (временной ряд, тренд, сезонная вариация, ошибки MAD и MSE).

47. Расчет сезонной вариации в аддитивных моделях. Центрированная скользящая средняя.

48. Десезонализация данных в аддитивных моделях.

49. Расчет уравнения тренда в аддитивных моделях.

50. Расчет ошибок в аддитивных моделях.

51. Прогнозирование в аддитивных моделях.

52. Расчет сезонной вариации в мультипликативных моделях.

53. Десезонализация данных в мультипликативных моделях.

54. Расчет уравнения тренда в мультипликативных моделях.

55. Расчет ошибок в мультипликативных моделях.

56. Прогнозирование в мультипликативных моделях.

57. Экспоненциальное сглаживание. Простая модель экспоненциального сглаживания.

58. Экспоненциальное сглаживание с поправкой на тренд.

59. Системы одновременных уравнений. Модель «спрос-предложение». Кейнсианская модель формирования доходов. Экзогенные и эндогенные переменные. Структурные уравнения модели. Приведенные уравнения.

60. Косвенный метод наименьших квадратов (КМНК).

61. Проблема идентификации. Идентифицируемая, неидентифицируемая и сверхидентифицируемая системы уравнений.

62. Необходимые условия идентификации.

63. Модель денежного рынка. Идентификация, оценка параметров.

64. Модифицированная модель Кейнса. Идентификация, оценка параметров.

65. Модифицированная модель «доход-потребление». Идентификация, оценка параметров.

66. Двухшаговый МНК (ДМНК).

ЛИТЕРАТУРА

Аронович А. Б., Афанасьев М. Ю., Суворов Б. П. Сборник задач по исследованию операций. — М.: Издательство МГУ, 1997.

Большаков А. С. Моделирование в менеджменте. — М.: Филинъ, 2000.

Бородич С. А. Эконометрика. — Мн.: Новое знание, 2001.

Ниворожкина Л. И. и др. Основы статистики с элементами теории вероятностей для экономистов. — Р-н/Д: Феникс, 1999.

Задания для контрольной работы по курсу «Эконометрика»

1–10. Автомат фасует сахар в пакеты. Проведена случайная выборка объемом n пакетов. Средний вес пакета сахара в выборке $\overline{X}$ кг, выборочное стандартное отклонение s кг. Найти доверительный интервал для среднего веса пакета сахара в генеральной совокупности с доверительной вероятностью p в случае:

а) стандартное отклонение автомата σ кг;

б) стандартное отклонение автомата неизвестно.

Определить необходимый объем выборки для достижения ширины доверительного интервала $\pm\Delta$. Проверить гипотезу о равенстве генеральной средней 1 кг.

Вариант	$\overline{X}$	n	σ	Δ	p	s
1	0,99	30	0,01	0,10	0,95	0,05
2	0,98	34	0,07	0,15	0,99	0,10
3	0,97	33	0,03	0,18	0,95	0,04
4	0,96	35	0,06	0,12	0,99	0,08
5	0,95	36	0,09	0,19	0,95	0,02
6	1,01	32	0,02	0,11	0,99	0,09
7	1,02	37	0,08	0,13	0,95	0,06
8	1,03	38	0,04	0,16	0,99	0,03
9	1,04	39	0,10	0,14	0,95	0,07
10	1,05	31	0,05	0,17	0,99	0,01

11–20. Проведена выборка объемом n_1 деталей. r_1 из них оказались бракованными. Найти доверительный интервал доли бракованных изделий в генеральной совокупности для доверительной вероятности p. Определить необходимый объем выборки для достижения ширины доверительного интервала $\pm\Delta$.

В повторной выборке объема n_2 r_2 деталей оказались бракованными. Понизилась ли доля брака?

Вариант	n_1	r_1	Δ	p	n_2	r_2
11	1000	200	0,01	0,95	1100	190
12	1100	190	0,02	0,99	1150	185
13	1200	180	0,09	0,95	1250	170
14	1300	170	0,08	0,99	1350	165
15	1400	160	0,07	0,95	1430	155

Вариант	n_1	r_1	Δ	p	n_2	r_2
16	1500	150	0,03	0,99	1570	140
17	1600	140	0,04	0,95	1620	135
18	1700	130	0,06	0,99	1780	120
19	1800	120	0,02	0,95	1900	115
20	1900	110	0,05	0,99	2000	108

21–30. Для производства каждой из n_1 деталей по первой технологии было затрачено в среднем $\overline{X}_1$ с (выборочная дисперсия s_1^2 с^2). Для производства каждой из n_2 деталей по второй технологии было затрачено в среднем $\overline{X}_2$ с (выборочная дисперсия s_2^2 с^2). Можно ли сделать вывод, что по первой технологии требуется в среднем больше времени для производства одной детали? Доверительная вероятность равна p.

Вариант	$\overline{X}_1$	s_1^2	$\overline{X}_2$	s_2^2	p
21	38	4	31	2	0,95
22	39	5	32	3	0,99
23	33	7	31	8	0,95
24	37	8	34	7	0,99
25	35	4	32	5	0,95
26	36	5	36	4	0,99
27	37	7	35	7	0,95
28	38	8	33	8	0,99
29	39	3	40	5	0,95
30	40	2	34	4	0,99

31–40. Проводились испытания нового лекарства. В эксперименте участвовали n_1 мужчин и n_2 женщин. У m_1 мужчин и m_2 женщин наблюдались побочные эффекты. Можно ли утверждать, что побочные эффекты от нового лекарства у женщин возникают реже, чем у мужчин? Доверительная вероятность равна p.

Вариант	n_1	m_1	p	n_2	m_2
31	1000	200	0,95	1100	190
32	1100	190	0,99	1150	185
33	1200	180	0,95	1250	170
34	1300	170	0,99	1350	165
35	1400	160	0,95	1430	155
36	1500	150	0,99	1570	140
37	1600	140	0,95	1620	135
38	1700	130	0,99	1780	120
39	1800	120	0,95	1900	115
40	1900	110	0,99	2000	108

41–50. По результатам наблюдений найти оценки коэффициентов уравнения линейной регрессии $y = \beta_0 + \beta_1 x$, коэффициенты корреляции и детерминации, проверить гипотезу о наличии линейной связи. Если гипотеза верна, то построить доверительные интервалы для коэффициентов теоретического уравнения линейной регрессии, для среднего значения результативного признака при данном x_0 и для индивидуальных значений результативного признака при данном x_0. Доверительная вероятность равна p.

Вариант	x					y					p	x_0
41	1	5	3	4	7	1	5	5	2	8	0,95	2
42	3	6	7	8	7	1	3	5	5	4	0,99	4
43	4	7	5	4	5	3	1	2	2	1	0,95	6
44	9	8	3	4	1	0	1	4	3	5	0,99	7
45	1	0	3	3	0	2	3	5	6	4	0,95	2
46	0	4	7	8	5	2	6	8	7	5	0,99	6
47	4	2	3	4	3	8	6	8	7	6	0,95	5
48	7	5	1	0	3	8	6	4	2	4	0,99	4
49	3	5	7	2	5	1	3	5	0	1	0,95	4
50	4	4	8	9	5	6	2	9	9	4	0,99	7

51–60. По результатам наблюдений найти точечные и интервальные оценки коэффициентов уравнения линейной регрессии $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ и проверить общее качество уравнения линейной регрессии. Все ли коэффициенты статистически значимы? Проверить наличие гетероскедастичности с помощью теста ранговой корреляции Спирмена. Определить наличие автокорреляции с помощью критерия Дарбина-Уотсона. Выяснить, есть ли в модели мультиколлинеарность? Доверительная вероятность равна 0,95.

	51	**52**	**53**	**54**	**55**	**56**	**57**	**58**	**59**	**60**
x_1	8	8	7	1	1	2	1	6	1	3
	2	6	1	8	4	5	9	9	8	1
	9	9	5	2	6	7	4	2	3	5
	7	6	5	5	9	1	1	5	2	4
	1	8	5	1	9	3	5	8	2	9
	8	3	2	4	1	1	2	2	4	9
	3	5	8	8	2	5	1	4	1	3
	3	8	2	5	2	1	4	5	9	1
	4	2	9	2	5	2	1	7	5	2
	1	8	6	1	8	2	3	7	4	8
x_2	1	2	8	8	3	3	9	5	2	5
	5	2	8	1	5	3	6	2	5	7

	51	52	53	54	55	56	57	58	59	60
	0	9	5	4	2	4	8	8	3	2
	5	9	9	7	1	6	2	7	4	1
	8	9	6	2	5	8	4	3	1	2
	8	3	2	7	2	1	3	5	2	1
	3	3	1	1	5	5	7	8	7	3
	1	2	9	8	8	7	2	4	6	7
	5	3	5	5	3	6	5	6	3	4
	2	7	4	6	7	2	2	7	8	8
y	2	8	1	5	4	1	6	4	3	7
	1	5	3	2	6	9	1	9	5	7
	5	3	5	8	8	4	9	1	4	5
	2	8	4	3	5	1	3	2	6	4
	4	9	9	8	3	5	2	6	9	2
	6	4	5	5	5	2	7	1	2	8
	6	1	1	9	1	7	1	2	1	1
	9	4	9	4	4	1	4	4	3	3
	7	6	1	1	1	3	3	6	9	4
	3	7	5	5	2	2	9	6	4	2

61–70. Два человека дегустируют 10 сортов кофе. Каждый из них расположил эти сорта в порядке убывания предпочтения. Есть ли какая-нибудь связь между этими результатами? Доверительная вероятность равна p.

	61	62	63	64	65	66	67	68	69	70
Дег.1	8	7	8	7	1	1	3	5	2	9
	10	6	2	1	4	9	7	1	1	3
	1	5	9	5	6	4	9	9	4	1
	2	9	7	10	9	10	6	10	10	4
	6	1	1	2	10	5	10	6	5	7
	3	2	10	8	2	2	4	2	3	2
	4	4	3	9	5	3	2	4	6	10
	5	8	4	6	8	6	5	7	8	5
	9	3	5	3	3	8	1	3	7	6
	7	10	6	4	7	7	8	8	9	8
Дег.2	9	8	8	1	2	6	7	5	3	6
	8	3	6	8	5	9	6	1	1	10
	1	2	9	2	7	2	4	4	7	4
	10	4	10	5	1	5	1	9	2	2
	2	1	3	10	3	8	10	10	5	1
	6	10	5	4	10	10	9	8	10	7
	7	9	2	7	6	4	8	3	4	8
	4	6	7	6	4	7	3	2	9	9
	3	5	1	3	8	3	2	6	6	5
	5	7	4	9	9	1	5	7	8	3
p	0,95	0,99	0,95	0,99	0,95	0,99	0,95	0,99	0,95	0,99

71–80. Дать прогноз объема продаж на следующие 3 дня.

Вариант	пн	вт	ср	чт	пт	сб	вск
71	1	3	2	9	2	8	5
	3	3	1	6	4	10	3
72	3	4	2	6	7	12	5
	1	3	2	7	3	6	9
73	9	4	7	5	4	2	3
	13	6	8	6	7	5	2
74	1	5	3	5	4	10	5
	2	3	2	7	5	9	4
75	1	5	2	6	2	9	8
	1	4	3	7	7	11	6
76	8	3	5	4	3	9	2
	9	7	8	8	5	4	6
77	2	6	4	6	7	9	10
	2	5	1	7	5	11	15
78	15	5	8	6	3	8	4
	10	6	9	6	5	6	6
79	1	3	4	7	3	6	9
	2	3	1	7	2	9	10
80	1	4	2	5	5	11	17
	2	7	9	6	4	9	7

81–90. Дать прогноз объема продаж на 11-ю неделю методами простого экспоненциального сглаживания и экспоненциального сглаживания с поправкой на тренд. Прогноз объема продаж на 1-ю неделю равен F_1. $T_1 = 0$.

Вариант	α	b	F_1
81	0,7	0,4	3
82	0,8	0,3	2
83	0,9	0,2	2
84	0,7	0,5	3
85	0,8	0,4	2
86	0,9	0,3	3
87	0,7	0,2	4
88	0,8	0,5	2
89	0,9	0,7	2
90	0,7	0,6	3

Содержание

ISBN 5-93840-079-1




ООО «Издательство РДЛ».
117334, Москва, ул. Вавилова, д. 30/6.
Тел.: (095) 135-98-93. e-mail: rdl@rinet.ru
Лицензия ИД № 00834 от 25 января 2000 г.

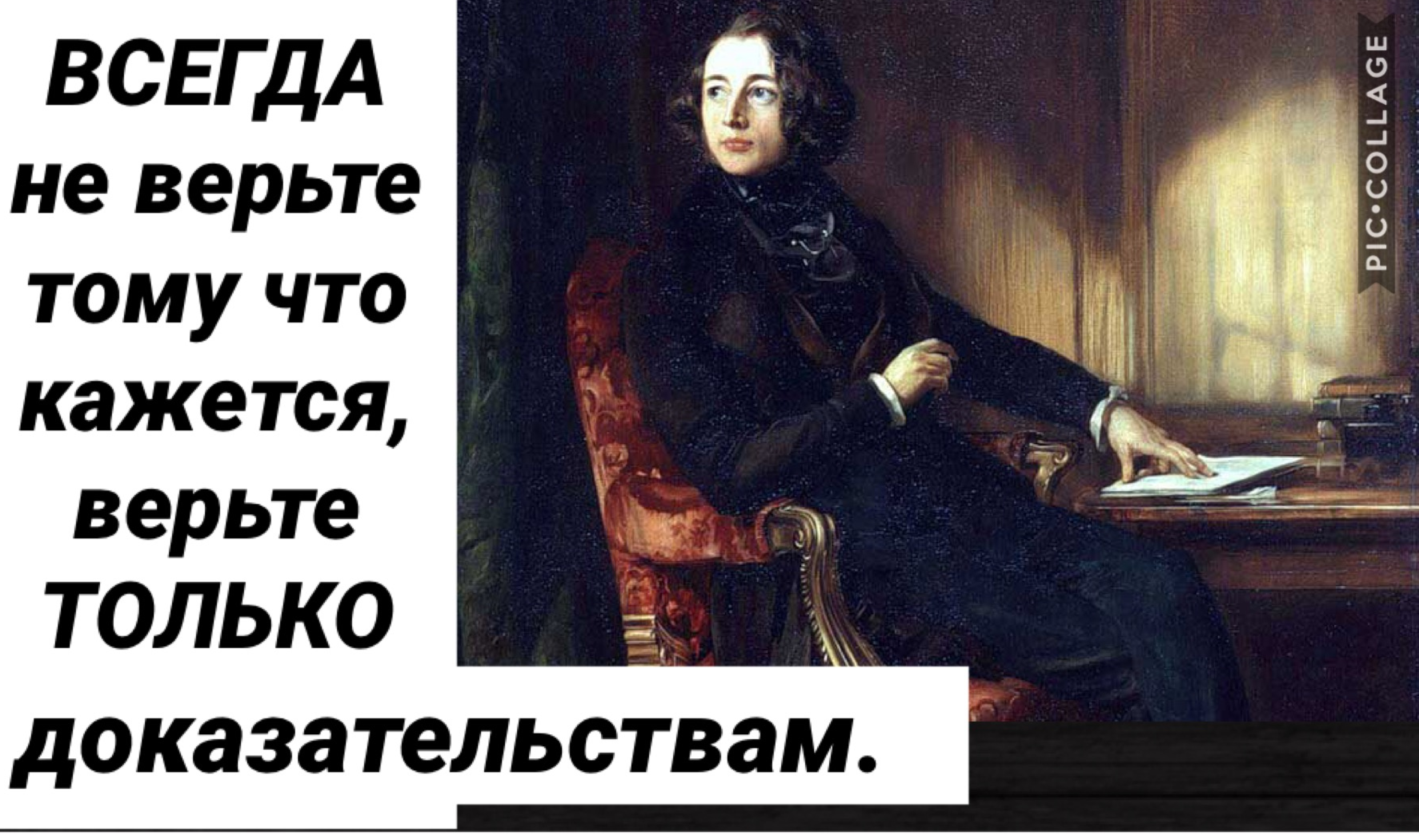
Сдано в набор 8.12.2004.
Подписано в печать 20.02.2005.
Формат 84x108 1/32. Гарнитура Школьная.
Печать офсетная. Тираж 700 экз.
Заказ № 339.

Начальник редакции *В. М. Дубильт.*
Научный редактор *В. М. Трояновский.*

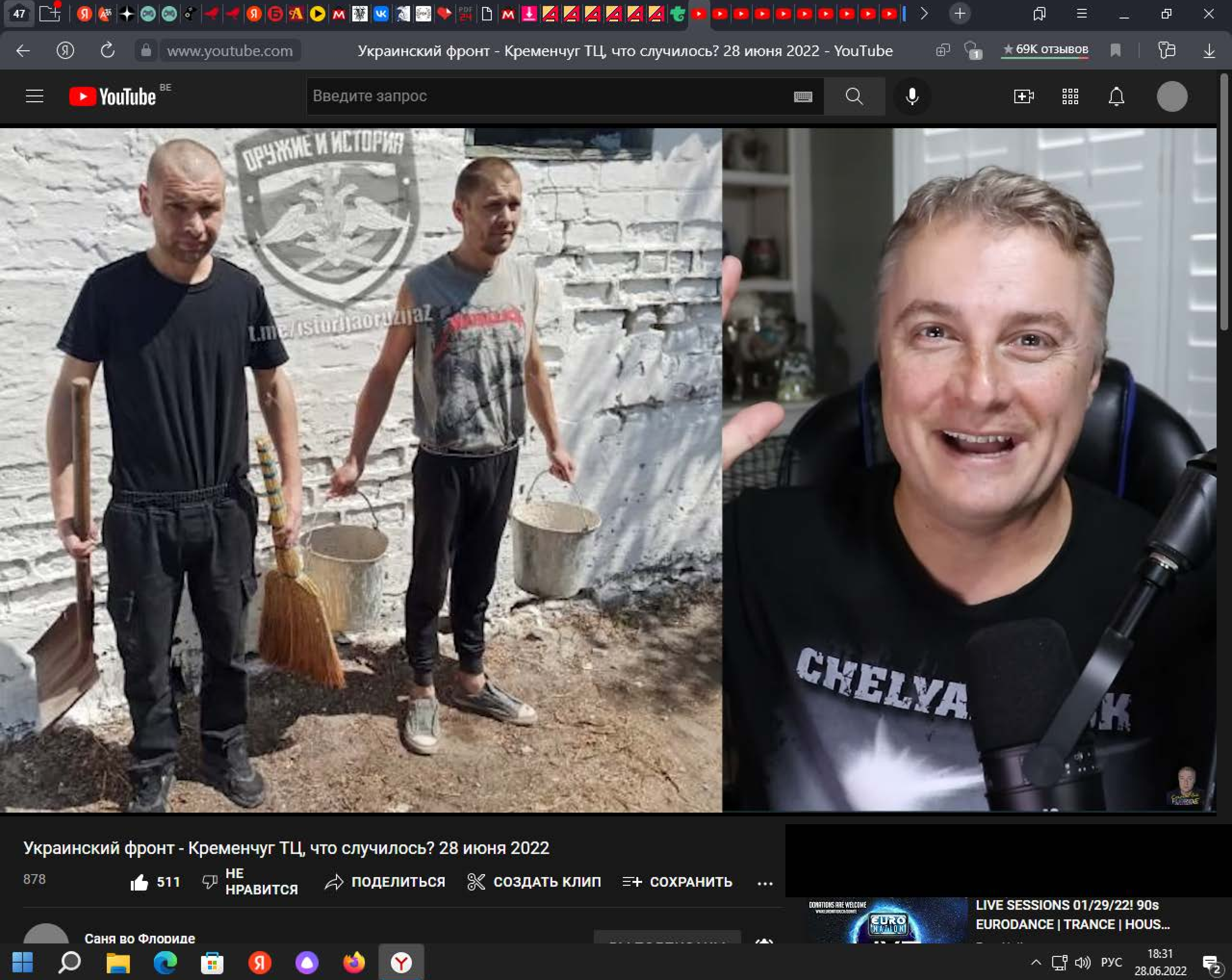
Отпечатано в Загорской типографии
141300, Московская область,
г. Сергиев Посад, пр. Красной Армии, д. 212Б.

Издательство РДЛ».
п. Вавилова, д. 30/6.
. e-mail: rdl@rinet.ru
от 25 января 2000 г.
но в набор 8.12.2004.
в печать 20.02.2005.
Гарнитура Школьная.
тная. Тираж 700 экз.
Заказ № 339.
акции В. М. Дубильт.
р В. М. Трояновский.
агорской типографии
Московская область,
асной Армии, д. 212Б.





Чарльз Диккенс. «Большие надежды» 1861 г.



www.youtube.com
Украинский фронт - Кременчуг ТЦ, что случилось? 28 июня 2022 - YouTube
69K отзывов
YouTube BE
Введите запрос
ОРУЖИЕ И ИСТОРИЯ
t.me/istorijaoruzijaZ
CHELYA
Украинский фронт - Кременчуг ТЦ, что случилось? 28 июня 2022
878
511
НЕ НРАВИТСЯ
ПОДЕЛИТЬСЯ
СОЗДАТЬ КЛИП
СОХРАНИТЬ
LIVE SESSIONS 01/29/22! 90s EURODANCE | TRANCE | HOUS...
Саня во Флориде
РУС
18:31
28.06.2022